DOCUMENT RESUME

ED 355 196 SP 034 282

AUTHOR	Estes, Gary D.; And Others
TITLE	Assessment Component of the California New Teacher Project: Second Year Report.
INSTITUTION	Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY	California State Dept. of Education, Sacramento.
PUB DATE	Jun 92
NOTE	606p.; For related documents, see ED 323 197, ED 342 761, and SP 034 278-281.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF03/PC25 Plus Postage.
DESCRIPTORS	Beginning Teachers; Elementary Secondary Education; English Instruction; *Evaluation Methods; *Evaluation Research; Evaluators; Higher Education; Language Arts; *Measures (Individuals); *Pilot Projects; Policy Formation; Sciences; Secondary School Teachers; Social Studies; *Teacher Certification; Teacher Education; Teacher Evaluation
IDENTIFIERS	*Assessments of Performance in Teaching; *California New Teacher Project; Pedagogical Content Knowledge; Reform Efforts

ABSTRACT

The California New Teacher Project (CNTP) commissioned pilot tests of assessment instruments during 1990. This document is the final report and analysis of the administration and scoring of these assessment instruments. The document, organized into 11 chapters, begins with an introduction describing research on new and experienced teachers, support and assessment of new teachers in California, California teacher credentialing reforms, and the CNTP. The next chapter describes the pilot test design and processes used to evaluate the assessment approaches. In the chapters that follow, each of the assessment instruments is described along with a discussion of ease of administration, scoring, content and format, costs, and technical qualities of the instrument. The instruments are presented in the following order: Structured Simulation Tasks for Secondary Life/General Science Teachers; Science Laboratory Assessment; Language Arts Pedagogical Knowledge Assessment; Structured Simulation Tasks for Secondary English Teachers; Secondary English Assessment: Assessment Center Activities; Secondary English Assessment: Portfolio Activity; Semi-Structured Interview in Secondary Social Studies; and Assessment of Competence in Monitoring Student Achievement in the Classroom. The report concludes with a summary of strengths and weaknesses of the assessment approaches represented by these instruments, conclusions about the effective design of training for assessors and/or scorers, and an identification or argumentation of policy issues beyond those discussed in the first-year report which will affect the design of a teacher assessment system. Eight appendices provide statistics on the assessment instruments; 61 tables and 12 figures are included. (LL)

SP

ED355196

ASSESSMENT COMPONENT OF THE CALIFORNIA NEW TEACHER PROJECT: SECOND YEAR REPORT

JUNE 1992

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Ross
____
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Far West Laboratory for Educational Research and Development
730 Harrison Street, San Francisco, CA 94107-1242 (415) 565-3000

SP034282

BEST COPY AVAILABLE

ASSESSMENT COMPONENT OF THE CALIFORNIA NEW TEACHER PROJECT:

SECOND YEAR REPORT

Far West Laboratory for
Educational Research and Development

Gary D. Estes
Kendyll Stansbury
Claudia Long
Kenneth Wolf
Gary Lichtenstein

June 1992

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Research on New and Experienced Teachers 1.2

Support and Assessment of New Teachers in California 1.3

California Teacher Credentialing Reforms 1.4

The California New Teacher Project 1.5

  Assessment Component of the California New Teacher Project 1.6

1990 Pilot Testing 1.10

CHAPTER 2: PILOT TEST DESIGN AND ANALYSIS

Design of Pilot Tests 2.1

  Sources of Instrumentation 2.1
  Sampling Plans 2.3
  Data Collection 2.4
  Data Reduction 2.6

Overview of Analytic Categories 2.7

  Administration of Assessment 2.7
  Assessment Content 2.7
  Assessment Format 2.8
  Cost Analysis 2.8
  Technical Quality 2.8

CHAPTER 3: STRUCTURED SIMULATION TASKS FOR SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Administration of Assessment 3.3

  Overview 3.3
  Logistics 3.5

    Identifying teacher samples 3.5
    Sending orientation materials 3.5

Security 3.6
Assessors and Their Training 3.7

Teacher and Assessor Impressions of Administration 3.7

Scoring 3.7

Scoring Process 3.7
Scorers and Their Training 3.10

Assessment Content 3.12

Congruence with California Model Curriculum Guides and Frameworks 3.13
Extent of Coverage of California Standards for Beginning Teachers 3.20
Job-Relatedness 3.24

Teacher perceptions 3.24
Scorer perceptions 3.26

Appropriateness for Beginning Teachers 3.27

Perceptions 3.27
Performance on assessment tasks 3.30

Appropriateness across Contexts 3.33

Grade level 3.33
Diverse students 3.34

Fairness across Groups of Teachers 3.37
Appropriateness as a Method of Assessment 3.38

Comparison with other assessments 3.40

Assessment Format 3.42

Format Features 3.42
Clarity of Preparatory Materials 3.42
Clarity of Task Instructions 3.44
Length of Tasks 3.47
Clarity of Scoring Criteria and Procedures 3.47

Cost Analysis 3.48

Administration and Scoring Costs Estimate 3.48
Development and Pilot Testing Costs 3.49

Technical Quality 3.49

Development 3.51

Reliability 3.51

Interrater agreements 3.51
Internal consistency of the tasks and assessment 3.52
Intercorrelations among tasks 3.54

Validity of Agreement Through Group Comparisons 3.54

Content validity 3.54

Conclusions and Recommendations 3.57

Administration of Assessment 3.57
Scoring 3.58
Assessment Content 3.59
Assessment Format 3.60
Summary 3.61

CHAPTER 4: SCIENCE LABORATORY ASSESSMENT

Administration of Assessment 4.5

Overview 4.5
Logistics 4.6

Recruiting and training observers 4.6
Identifying the teacher sample 4.6
Scheduling the observations 4.8
Sending orientation materials 4.8
Collecting evaluation feedback 4.8

Security 4.8
Assessors and Their Training 4.9

Characteristics of the assessors 4.9
Training 4.9
Perceptions of training 4.10

Scoring 4.11
Teacher, Assessor, and FWL Staff Perceptions of Administration 4.12

Assessment Content 4.14

Congruence with the 1990 California Science Framework 4.15

Extent of Coverage of California Standards for Beginning Teachers 4.17
Job Relatedness 4.22
Appropriateness for Beginning Teachers 4.22

Perceptions 4.24
Performance on assessment 4.24

Appropriateness across Contexts 4.30

Across grade levels 4.30
Diverse students 4.32

Fairness across Groups of Teachers 4.35
Appropriateness as a Method of Assessment 4.35

Appropriateness 4.35
Comparison 4.36

Assessment Format 4.37

Clarity of the Teachers' Preparation Materials 4.38
Clarity of the Conference Questions 4.40
Clarity of the Forms and Process for Documentation and Rating 4.41

Guided note-taking form 4.44
Documentation sorting record 4.46
Summary report form 4.48

Cost Analysis 4.52

Administration and Scoring Cost Estimates 4.52

Assessor time and costs 4.52
Training costs for assessors 4.53
Other costs 4.53

  - Development and Pilot Testing Costs 4.53
  - Cost Summary 4.54

- Technical Quality 4.54
  - Development 4.54
  - Reliability 4.56
  - Validity 4.57

- Conclusions and Recommendations 4.57
  - Administration of Assessment 4.57
  - Assessment Content 4.59
  - Assessment Format 4.61
  - Summary 4.62

CHAPTER 5: LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT

- Administration of Assessments 5.4
  - Overview 5.4
  - Logistics 5.6
    - Identifying teacher samples 5.6
    - Sending orientation materials 5.6
    - Assessment administration 5.6
    - Collecting evaluation feedback 5.7
  - Security 5.7

- Scoring 5.8
  - Scoring Process 5.8
  - Scorers and Their Training 5.9
    - Scoring characteristics 5.9
    - Training 5.9
    - Perceptions of training 5.10
  - Teacher and FWL Staff Impressions of Administration 5.11

- Assessment Content 5.11
  - Congruence with California Curriculum Guides and Frameworks 5.12
  - Extent of Coverage of California Standards for Beginning Teachers 5.16

Job-Relatedness . . . 5.19

Appropriateness for Beginning Teachers . . . 5.19

- Perceptions . . . 5.19
- Performance on assessment . . . 5.22

Appropriateness across Contexts . . . 5.24

- Grade level . . . 5.24
- Diverse students . . . 5.24

Fairness across Groups of Teachers . . . 5.27

Appropriateness as a Method of Assessment . . . 5.28

Comparison with other assessments . . . 5.29

Assessment Format . . . 5.30

- Clarity of Teacher Preparation Materials . . . 5.30
- Clarity of Task Materials . . . 5.31
  - Suggestions for improving the task directions . . . 5.31
  - Suggestions for improving the questions . . . 5.32
  - Other suggestions for improving the materials . . . 5.32
- Clarity of the Scoring Criteria and Procedures . . . 5.33

Cost Analysis . . . 5.36

- Administration and Scoring Cost Estimates . . . 5.36
- Development and Pilot Testing Costs . . . 5.37

Technical Quality . . . 5.37

- Development . . . 5.37
- Reliability . . . 5.39
  - Interrater agreements . . . 5.39
  - Interrater correlations . . . 5.41
- Validity of Agreement Through Group Comparisons . . . 5.42
- Content Validity . . . 5.42

Conclusions and Recommendations 5.43

Administration of Assessment 5.43
Assessment Content 5.45
Assessment Format 5.47
Summary 5.49

CHAPTER 6: STRUCTURED SIMULATION TASKS FOR SECONDARY ENGLISH TEACHERS

The Structured Simulation Tasks 6.1

Administration of Assessment Tasks 6.3

Overview 6.3
Logistics 6.4

Identifying teacher samples 6.4
Sending orientation materials 6.4
Administering the tasks 6.4
Collecting evaluation feedback 6.5

The Teacher Sample 6.5

Background characteristics and preparation 6.5
Teaching contexts 6.5

Security 6.5
Assessors and Their Training 6.7
Teacher and Assessor Impressionsof Administration Logistics 6.7

Scoring 6.7

Logistics 6.7
Scorers and Their Training 6.8
Scoring Process 6.8

Structured Simulation Tasks' Content 6.12

Coverage of California English/Language Arts Framework 6.12
Extent of Coverage of California Standards for Beginning Teachers 6.18
Job-relatedness 6.24

Teacher perceptions 6.24
Scorer perceptions 6.25

- Appropriateness for Beginning Teachers 6.25
  - Scorer perceptions 6.28
- Performance on Structured Simulation Tasks 6.30
  - Analytic versus holistic scoring 6.34
  - Determining performance standards 6.42
- Appropriateness across Contexts 6.42
- Fairness across Groups of Teachers 6.44
- Appropriateness as a Method of Assessment 6.45

Structured Simulation Tasks' Format 6.46

- Clarity of Preparatory Materials 6.46
- Clarity of Task Instructions 6.46

Cost Analysis 6.49

- Administration and Scoring Cost Estimates 6.49
- Development and Pilot Testing Costs 6.50

Technical Quality 6.50

- Development 6.50

Reliability 6.53

Inter-correlations Among Tasks 6.55

Validity 6.57

Summary 6.59

Recommendations 6.62

- Administration 6.62
- Development 6.62

Conclusion 6.65

CHAPTER 7: SECONDARY ENGLISH ASSESSMENT: ASSESSMENT CENTER ACTIVITIES

Administration of Assessment Center Activities 7.4

  - Overview 7.4
  - Logistics 7.5
    - Identifying the teacher sample 7.5
    - Recruiting and training of assessors 7.7
    - Scheduling/Arranging the assessments 7.7
    - Developing and sending the orientation materials 7.7
    - Collecting evaluation feedback 7.8
  - Security 7.8
  - Assessors and Their Training 7.9
    - Characteristics of the assessors 7.9
    - Training 7.9
    - Perceptions of training 7.10
  - Scoring 7.11
  - Teacher, Assessor, and FWL Staff Perceptions of Administration 7.12
- Assessment Content 7.13
  - Congruence with the California English/Language Arts Framework and Handbooks 7.14
  - Extent of Coverage of California Standards for Beginning Teachers 7.17
  - Job-relatedness 7.19
  - Appropriateness for Beginning Teachers 7.22
    - Perceptions 7.22
    - Performance on assessment 7.25
  - Appropriateness across Contexts 7.27
    - Grade level 7.28
    - Diverse students 7.28
  - Fairness across Groups of Teachers 7.30
  - Appropriateness as a Method of Assessment 7.32
    - Appropriateness 7.33
    - Comparison of activities with other assessments 7.34
- Assessment Format 7.36

Clarity of the Teachers' Preparation Materials 7.36
Appropriateness of Time Allotted for Each Activity 7.40
Clarity of the Rating Forms and Process 7.43

Cost Analysis 7.51

Administration and Scoring Cost Estimates 7.51
Development and Pilot Testing Costs 7.52

Technical Quality 7.52

Development 7.52
Reliability 7.54

Interrater agreements 7.54
Interrater correlations 7.56
Internal consistency of the tasks and assessment 7.58
Intercorrelations among activities 7.59

Validity of Agreement Through Group Comparisons 7.59

Content validity 7.62

Conclusions and Recommendations 7.62

Administration of Assessment 7.62
Assessment Content 7.63
Assessment Format 7.66

Summary 7.68

CHAPTER 8: SECONDARY ENGLISH ASSESSMENT: PORTFOLIO ACTIVITY

Administration of Portfolio Activity 8.2

Overview 8.4
Logistics 8.4

Contacting the Identified Teachers 8.4
Making the Follow-up Phone Calls 8.6
Arranging for the Mailing of Portfolios 8.6
Recruiting and Training of Scorers 8.6
Collecting Evaluation Feedback 8.6

- Security 8.6
- Assessors and Their Training 8.7
- Teacher and FWL Staff Perceptions of Administration 8.7
- Scoring 8.9
  - Scoring Process 8.9
  - Characteristics of the Scorers 8.9
  - Training 8.10
  - Perceptions of Training 8.10

Assessment Content 8.11

- Congruence With the California English/Language Arts Framework and Handbook 8.11
- Extent of Coverage of California Standards for Beginning Teachers 8.12
- Job-relatedness 8.16
- Appropriateness for Beginning Teachers 8.16
  - Perceptions 8.16
  - Performance on Assessment 8.20
- Appropriateness across Contexts 8.23
  - Grade level 8.23
  - Diverse students 8.24
- Fairness across Groups of Teachers 8.26
- Appropriateness as a Method of Assessment 8.26

Assessment Format 8.28

- The Construction of the Portfolio 8.29
- Clarity of the Scoring System and Response Form 8.32

Cost Analysis 8.36

- Administration and Scoring Costs Estimate 8.37
- Development and Pilot Testing Costs 8.39

Technical Quality 8.39

- Development 8.41
- Reliability 8.41

Interrater agreements 8.41
Internal consistency of the tasks and assessment 8.41
Intercorrelations among activities 8.44

Validity of Agreement Through Group Comparisons 8.44

Content validity 8.47

Conclusions and Recommendations 8.47

Administration of Assessment 8.47
Assessmentent Content 8.49
Assessment Format 8.51

Construction of the Portfolio 8.51
Scoring Process and Response Form 8.52

Summary 8.53

CHAPTER 9: SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES

Administration of Assessment 9.2

Overview 9.2
Logistics 9.4

Identifying Teacher Samples 9.4
Orientation to the Assessment 9.5

Security 9.5
Assessors and Their Training 9.5
Teacher and Assessor Impressions of Administration 9.7

Scoring 9.7

Scoring Process 9.7

Assessment Content 9.9

Congruence With California Model Curriculum Guides and Frameworks 9.10
Extent of Coverage of California Standards for Beginning Teachers 9.13
Job-relatedness 9.17

Teacher Perceptions 9.17
Assessor Perceptions 9.18

Appropriateness for Beginning Teachers 9.19

Appropriateness across Contexts 9.21

Grade level 9.21
Diverse students 9.21

Fairness across Groups of Teachers 9.24
Appropriateness as a Method of Assessment 9.25
Comparisons With Other Assessments 9.27

Assessment Format 9.28

Format Features 9.29
Clarity of Preparatory Material 9.29
Clarity of Task Instructions 9.30
Length of Tasks 9.32
Success in Duplicating the Methodology in Another Subject Area 9.32

Cost Analysis 9.34

Technical Quality 9.34

Conclusions and Recommendations 9.34

Administration of Assessment 9.35
Scoring 9.36
Assessmentent Content 9.36
Assessment Format 9.37
Summary 9.39

CHAPTER 10: ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Administration of Assessments 10.3

Overview 10.3
Logistics 10.3

Identifying teacher samples 10.6
Orientation materials 10.6
Conducting the assessment 10.6

Conducting staff development 10.6
Obtaining feedback from the teachers 10.6

Security 10.7
Assessors and Their Training 10.7
Teacher Impressions of Administration 10.7

Scoring 10.8

Scoring Process 10.8
Scorers and Their Training 10.9

Scorers 10.9
Training of scorers 10.10
Perceptions of training 10.11

Assessment Content 10.11

Congruence with California Model Curriculum Guides and Frameworks 10.12

Language arts 10.13
Science 10.14
Social science 10.15
Mathematics 10.16

Extent of Coverage of California Standards for Beginning Teachers 10.17
Job-Relatedness 10.18

Teacher perceptions 10.18
Scorer perceptions 10.19

Appropriateness for Beginning Teachers 10.19

Teacher perceptions 10.19
Scorer perceptions 10.22
Performance on assessment tasks 10.22

Appropriateness across Contexts 10.24

Grade level and subject matter 10.24
Diverse students 10.26

Fairness across Groups of Teachers 10.29
Appropriateness as a Method of Assessment 10.29

Teacher perceptions 10.30
Scorer perceptions 10.30
Comparison with other assessments 10.31

Assessment Format 10.32

Format Features 10.32
Clarity of Assessment 10.32

Clarity of questions 10.32
Clarity of scoring criteria 10.34
Evaluation of staff development training 10.35

Cost Analysis 10.36

Administration and Scoring Costs 10.36
Development and Pilot Testing Costs 10.37

Technical Quality 10.39

Development 10.39
Reliability 10.40

Interrater agreements 10.40
Interrater correlations 10.40
Internal consistency of the assessment forms 10.40

Validity of Agreement through Group Comparisons 10.43

Content validity 10.44

Conclusions and Recommendations 10.46

Administration of Assessment 10.46
Scoring 10.46
Assessment Content 10.47
Assessment Format 10.49
Summary 10.49

CHAPTER 11: CONCLUSIONS

Assessment Approaches 11.1

Structured Simulation Tasks 11.1

Definition 11.1
Characteristics of instruments piloted 11.1
Strengths and weaknesses 11.3

Classroom Observations (Subject Matter Focus) 11.4

Definition 11.4
Characteristics of instrument piloted 11.4
Strengths and weaknesses 11.5

Semi-Structured Interviews 11.6

Definition 11.6
Characteristics of instrument pioleted 11.6
Strengths and weaknesses 11.6

Videotaped Teaching Episodes 11.7

Definition 11.7
Characteristics of instrument piloted 11.7
Strengths and weaknesses 11.8

Performance-Based Assessment Center Exercises 11.9

Definition 11.9
Characteristics of instrument piloted 11.9
Strengths and weaknesses 11.10

Portfolio 11.11

Definition 11.11
Characteristics of instrument piloted 11.11
Strengths and weaknesses 11.11

Guidelines for the Design of Training 11.13

Cost Estimates 11.14

Policy Issues 11.15

TABLES

TABLE 3.1 Pilot Test Participants: Secondary Life/General Science Teacher Assessment 3.4

TABLE 3.2 Coverage of the California Science Framework by the Secondary Life/General Science Teacher Assessment 3.14

TABLE 3.3 Extent of Coverage by the Secondary Life/General Life Science Teacher Assessment of California Standards for Beginning Teachers 3.25

TABLE 3.4 Teacher Performance, by Subpart, on the Secondary Life/General Science Teacher Assessment 3.31

TABLE 3.5 Teacher Perceptions of the Preparatory Materials for the Secondary Life/General Science Teacher Assessment 3.43

TABLE 3.6 Teacher Perceptions of the Clarity of Task Instructions for the Secondary Life/General Science Teacher Assessment 3.45

TABLE 3.7 Developmental and Pilot Test Costs for the Secondary Life/General Science Teacher Assessment 3.50

TABLE 3.8 Internal Consistency of Tasks: Secondary Life/General Science Teacher Assessment 3.53

TABLE 3.9 Intercorrelations Among Tasks: Secondary Life/General Science Teacher Assessment 3.55

TABLE 3.10 Trends of Mean Differences in Task Performance Between Candidates with Different Characteristics: Secondary Life/General Science Teacher Assessment 3.56

TABLE 4.1 Pilot Test Participants: Science Laboratory Assessment 4.7

TABLE 4.2 Coverage of the California Science Framework by the Science Laboratory Assessment 4.16

TABLE 4.3 Extent of Coverage by the Science Laboratory Assessment of California Standards for Beginning Teachers 4.23

TABLE 4.4 Suggested Changes to Pre-Observation Conference Questions: Science Laboratory Assessment 4.42

TABLE 4.5 Developmental and Pilot Test Costs for the Science Laboratory Assessment 4.55

TABLE 5.1 Some Characteristics of the Four LAPKA Scenarios for Assessing a Teacher's Pedagogical Content Knowledge in Language Arts 5.2

TABLE 5.2 Pilot Test Participants: Language Arts Pedagogical Knowledge Assessment (LAPKA) 5.5

TABLE 5.3 Congruence of the Language Arts Pedagogical Knowledge Assessment (LAPKA) with the English-Language Arts Model Curriculum Guide for Kindergarten through Grade Eight 5.15

TABLE 5.4 Extent of Coverage by the Language Arts Pedagogical Knowledge Assessment (LAPKA) of the California Standards for Beginning Teachers 5.20

TABLE 5.5 Performance Data for Pilot Test Teachers (N=42) for the Language Arts Pedagogical Knowledge Assessment (LAPKA) 5.23

TABLE 5.6 Developmental and Pilot Test Costs for the Language Arts Pedagogical Knowledge Assessment (LAPKA) 5.38

TABLE 6.1 Pilot Test Participants: Structured Simulation Tasks for Secondary English Teachers 6.6

TABLE 6.2 Coverage of English/Language Arts Framework 6.13

TABLE 6.3 Congruence with California Standards for Beginning Teachers 6.19

TABLE 6.4 Range of Scores, Mean, and Pass Rates for Tasks for Group I 6.32

TABLE 6.5 Range of Scores, Mean, and Pass Rates for Tasks for Group II 6.33

TABLE 6.6 Pilot Test Costs for Structured Simulation Tasks for Secondary English 6.51

TABLE 6.7 Cronbach Alpha Coefficients for the Six Tasks 6.54

TABLE 6.8 Correlations Among Tasks 6.56

TABLE 7.1 Pilot Test Participants: Secondary English Assessment 7.6

TABLE 7.2 Congruence of the Secondary English Assessment with the English-Language Arts Framework and Handbooks 7.16

TABLE 7.3 Extent of Coverage by the Secondary English Assessment of California Standards for Beginning Teachers 7.20

TABLE 7.4 The Number of Teachers Receiving Each Rating in the Evaluation Categories for Each Activity 7.26

TABLE 7.5 Developmental and Pilot Test Costs for the Secondary English Assessment 7.53

TABLE 7.6 Correlations Between Raters for the Secondary English Assessment Activities for Holistic Rating (RT) and Summed Ratings (S) 7.57

TABLE 7.7 Trends of Mean Differences Between Candidates with Different Characteristics for Activities and Evaluation Categories 7.60

TABLE 8.1 Pilot Test Participants: Portfolio Activity: Secondary English Assessment 8.5

TABLE 8.2 Congruence of the Secondary English Assessment's Portfolio Activity with the English/Language Arts Framework and the Literature Handbook 8.13

TABLE 8.3 Extent of Coverage by the Secondary English Assessment Portfolio Activity of the California Standards for Beginning Teachers 8.17

TABLE 8.4 Number and Percent of Teachers Receiving Each Rating in the Six Evaluation Categories of the Portfolio Activity 8.21

TABLE 8.5 Pilot Test Costs for the Portfolio Activity 8.40

TABLE 8.6 Internal Consistency of the Portfolio Activity and Its Parts 8.43

TABLE 8.7 Intercorrelations Among Parts of the Portfolio Activity 8.45

TABLE 8.8 Trends of Mean Differences Between Candidates With Different Characteristics for Parts of Portfolio Activity 8.46

TABLE 9.1 Pilot Test Participants: Semi-Structured Interview in Secondary Social Studies 9.3

TABLE 9.2 Coverage of the California History/Social Science Framework by the Semi-Structured Interview in Secondary Social Studies 9.11

TABLE 9.3 Extent of Coverage by the Semi-Structured Interview in Secondary Social Studies of California Standards for Beginning Teachers 9.15

TABLE 9.4 Teacher Perception of the Clarity of Instruction for the Semi-Structured Interview in Secondary Social Studies 9.31

TABLE 10.1 Participation in Staff Development Activities by Participation in Pre- or Post-Tests Assessment of Competence in Monitoring Student Achievement in the Classroom 10.4

TABLE 10.2 Pilot Test Participants: Assessment of Competence in Monitoring Student Achievement in the Classroom 10.5

TABLE 10.3 Topics of Exercises Reported Being Too Difficult: Assessment of Competence in Monitoring Student Achievement in the Classroom 10.21

TABLE 10.4 Teacher Performance by Form, Pre- or Post-Test, and Participation: Assessment of Competence in Monitoring Student Achievement in the Classroom 10.23

TABLE 10.5 Distribution of Scenarios in Exercises Across Grade Levels: Assessment of Competence in Monitoring Student Achievement in the Classroom 10.25

TABLE 10.6 Developmental and Pilot Test Costs for the Assessment of Competence in Monitoring Student Achievement in the Classroom 10.38

TABLE 10.7 Correlations Between Paired Ratings for the Assessment of Competence in Monitoring Student Achievement in the Classroom 10.42

TABLE 10.8 Trends of Mean Differences in Performance Between Candidates with Different Characteristics: Assessment of Competence in Monitoring Student Achievement in the Classroom 10.45

TABLE Statistical Comparison of Teacher Performance on the Secondary Life/General Science Teacher Assessment A.1

TABLE Statistical Comparison of Teacher Performance on the Language Arts Pedagogical Knowledge Assessment D.1

TABLE Statistical Comparison of Teacher Performance on the Secondary English Assessment E.1

TABLE Statistical Comparison of Teacher Performance on the Secondary English Assessment: Portfolio Activity F.1

TABLE Statistical Comparison of Group Performance on the Assessment of Competence in Monitoring Student Achievement G.1

FIGURES AND CHARTS

FIGURE 4.1 List of Domains and Elements: Science Laboratory Assessment 4.2

FIGURE 4.2 Three Elements and Defining Indicators of the Materials/ Equipment Domain: Science Laboratory Assessment 4.3

FIGURE 4.3 Percent of Teachers Receiving a "2" Rating on Each Domain 4.25

FIGURE 4.4 Domains Teachers Believe Could Only Be Passed W/ 2 Yrs. Classroom Experience 4.28

FIGURE 4.5 Domains Teachers Believe Could Be Passed Immediately After Student Teaching 4.29

FIGURE 4.6 Percent of Teachers By Grade Level Receiving a "Two" Rating on Each Domain 4.31

FIGURE 5.1 Percent Agreement of Raters for the Language Arts Pedagogical Knowledge Assessment 5.40

FIGURE 7.1 Response Form A: Responding to Student Writing 7.3

FIGURE 7.2 Percent Agreement Between Raters for the Secondary English Assessment Activities 7.55

CHART 8.1 Responce Form D: Evaluation of General Pedagogical Abilities 8.3

FIGURE 8.1 Percent Agreement Rates for the Portfolio Activity 8.42

FIGURE 10.1 Rating Differences Between Scorers 10.41

APPENDICES

APPENDIX A: Statistical Comparison of Teacher Performance on the Secondary Life/General Science Teacher Assessment A.1

APPENDIX B: Science Laboratory Assessment Content and Forms B.1

APPENDIX C: An Example of a Scoring Sheet for the Language Arts Pedagogical Knowledge Assessment C.1

APPENDIX D: Statistical Comparison of Teacher Performance on the Language Language Arts Pedagogical Knowledge Assessment D.1

APPENDIX E: Statistical Comparison of Teacher Performance on the Secondary English Assessment E.1

APPENDIX F: Statistical Comparison of Teacher Performance on the Secondary English Assessment: Portfolio Activities F.1

APPENDIX G: Statistical Comparison of Group Performance on the Assessment of Competence in Monitoring Student Achievement G.1

CHAPTER 1:

INTRODUCTION

Recent analyses of education in the United States have identified significant areas of ineffectiveness (Boyer, 1983; Goodlad, 1984; President's Commission for Excellence in Education, 1983), and have included important proposals for reform (Holmes Group, 1986; Shulman, 1987; Carnegie Corporation, 1986). Many of these analyses and proposals have addressed the quality of the teaching force, with particular focus on the preparation, support, and credentialing of new teachers. Some of the leading educational scholars in the nation have concluded that the standards for membership in the teaching profession are inadequate, that few states require beginning teachers to attain acceptable levels of competence in classroom teaching, and that the quality of instruction in the public schools suffers as a result of fragmented approaches to teacher preparation, certification, induction and career development.

In the growing literature on educational reform, the emphasis on new teachers has been part of a broader discussion of promoting teaching as a profession (e.g., Wise and Darling-Hammond, 1987; Shulman and Sykes, 1986). Several leading advocates of educational reform have examined the standards of other professions (medicine, law, architecture, engineering, accountancy, etc.), and have argued that more rigorous and comprehensive assessments of teachers' knowledge and competence should be developed and adopted (Holmes Group, 1986; Shulman, 1987; Carnegie Corporation, 1986). After examining the histories of several professions, these researchers have suggested that the stature of a profession depends in part on the extent to which it verifies the professional knowledge and competence of each member in a broad assessment that must be passed in order to practice the profession in each state. In a series of reports that have been widely acclaimed, the critics recommended the development of stronger assessments as a way of strengthening instruction in the schools as well as public confidence in the teaching profession.

These same reform advocates have also recommended the creation of stronger support systems for beginning teachers in the schools. The practice of giving the most difficult teaching assignments to new teachers is increasingly identified as a major cause of the high rate at which novices leave teaching (Griffin and Millies, 1986; Ward, 1991; Odell, 1986; Hurling-Austin, 1986; Ryan, 1980). Similarly, the historically weak systems of supervision, evaluation, and assistance for beginning teachers have been based on a mistaken assumption that the completion of teacher preparation programs in universities provides a "complete" basis for the successful practice of school teaching. In fact, prospective teachers need structures to further their preparation in ways that bring together the elements of undergraduate preparation, teacher education, student teaching, and initial classroom work in a comprehensive way. This need is particularly vital in California in the 1990's where the diversity of backgrounds, languages, and academic preparation make teaching especially challenging. Recent analyses of these conditions have suggested that comprehensive **support** and **assessment systems** must be added to the new teacher preparation and credentialing process in order to promote the successful induction of teachers into an effective, and respected profession.

Research on New and Experienced Teachers

The educational reform efforts in California and across the nation have been motivated, in part, by the literature which identifies the technical, socioemotional, and institutional needs of new teachers, and explores the differences between new and experienced teachers. New teachers, for example, report significant difficulties with the technical aspects of teaching, including classroom management (Veenman, 1984), curriculum implementation (Grant and Zeichner, 1981; Veenman, 1984; Berliner et al., 1987), and managing diversity within the classroom (Grant and Zeichner, 1981; Veenman, 1984; Borko et al., 1986; Berliner et al., 1987; Berliner et al., 1988).

Socioemotionally, many new teachers experience insecurity, self-doubts, and substantial stress as they face the problems of acquiring and developing materials, lesson plans and tests without the expertise and materials that seasoned teachers draw upon. The typically brief period of supervised practice before assuming full teaching responsibilities, combined with working conditions which usually isolate teachers from their peers, provide new teachers with only limited opportunities to develop realistic standards for their performance (Moir, 1990). Not surprisingly, new teachers both need and usually appreciate someone who is willing to listen to their problems -- both personal and professional -- and offer supportive and useful feedback (Borko et al., 1986).

Institutionally, new teachers face the problems of having to quickly become familiar with district and school policies, practices, and procedures; learning about resources and how to access them; and becoming integrated into the community of teachers in the school. Many new teachers experience difficulties and frustration in locating and absorbing this critical information (Grant and Zeichner, 1981; Odell, 1986).

The research on new teachers also focuses on identifying stages at which different skills develop. The knowledge base of teaching is very complex, and the period of training is brief--especially compared to other professions which tend to provide for a more gradual assumption of professional responsibilities (Wise and Darling-Hammond, 1987). Preservice courses and experiences, no matter how well structured, cannot fully prepare teacher candidates to perform as excellent practitioners in the classroom. The emerging literature on differences between new and experienced teachers suggests that some skills may be present in only rudimentary form in new teachers. Compared to new teachers, for example, experienced teachers are more likely to see lessons as composed of general pedagogical routines for specific purposes, such as introducing new concepts, applying concepts previously learned, reviewing content previously learned, collecting homework, etc. (Leinhardt, 1989). Expert teachers also see the subject matter organized in frameworks, while novice teachers see it as more of a collection of facts (Wilson, 1988; Leinhardt, 1989). Knowledge of students and student learning also seems to be a skill that develops with experience in teaching (Leinhardt, 1983; Wilson, 1988).

Support and Assessment of New Teachers in California

Becoming a teacher in California is much the same as in other states. An individual can qualify for a teaching position by earning a baccalaureate degree in any field, completing a one-year post-graduate program of teacher education, and passing standardized tests of basic skills and content knowledge. During the 1980's, the support systems for beginning teachers in California consisted largely of (1) cooperating classroom teachers who supervised candidates during student teaching, (2) mentor teachers who assist new teachers and train experienced colleagues, and (3) principals of school in which new teachers were hired. Similarly, the beginning teacher assessments consisted of standardized multiple-choice tests and the evaluation of performance during student teaching and probationary employment.

In 1984 and 1985, the traditional systems of new teacher support and assessment were examined in considerable depth in California (Commission on Teacher Credentialing, 1985; Commission on Teacher Quality, 1983; California Commission on the Teaching Profession, 1985). In 1987, the California Department of Education (CDE) and the Commission on Teacher Credentialing (CTC) cosponsored a series of policy seminars at Stanford University on "New Teachers For California: Issues of Support and Assessment." These and other analyses made the following conclusions: (1) Student teaching and the professional education courses that accompany student teaching are important elements of teacher preparation, but they are insufficient for many new teachers to become skillful, proficient professionals; (2) Student teachers practice in environments that are considerably different from the settings in which they ultimately teach; (3) In addition to state reviews of teacher education programs, candidates need to demonstrate their individual readiness for teaching through a candidate-based assessment system; (4) Many of the complexities and nuances of effective teaching are learned during the teacher's initial classroom work; and, finally, (5) New approaches to teacher credentialing should include a model of licensure that takes into account new teachers' classroom pedagogy, subject matter knowledge, and ability to relate to students.

Traditionally, the primary supervisors of new teachers have been site administrators. The growing demands on school principals have made it increasingly difficult, however, for these local educational leaders to attend to the needs of beginning teachers in timely, intensive ways. New teachers are most often employed in schools with large, crowded classrooms of students who are increasingly diverse in their languages, academic, and cultural backgrounds. Most principals don't have sufficient time (and in some cases, expertise) to provide high quality support for new teachers in these contexts. Experienced teachers, another logical source of support for new teachers, also lack sufficient time to provide intensive support. Moreover, to be effective supporters of new teachers, research suggests that experienced teachers need not only time, but also authority, compensation and training.

On the other side of the new teacher support coin is accountability. In 1984-85, the "reform commissions" concluded that the traditional assessments of beginning teachers were inadequate to the challenge of verifying the competence of each new professional. With little prior training, thousands of classroom teachers assessed the performances of student teachers on the basis of standards and criteria that were varied, unclear, and poorly related to the changing realities of California classrooms. Furthermore, knowing that the prospective teacher's career depended largely on the evaluation he or she received as a student teacher, virtually all supervisors awarded outstanding grades to the novices whom they also provided guidance and assistance. Evaluations made by school principals of their new hires were also rarely negative, and, due to the large differences between school districts, the standards and procedures used for evaluations tended to be uneven and unreliable.

California Teacher Credentialing Reforms

Several reform initiatives undertaken by the CTC and the CDE since 1985 have been devoted to the successful resolution of issues related to the comprehensive support and assessment of beginning teachers. In concert with local teachers and administrators, the CTC examined the tests that new teachers were required to pass for California teaching credentials. The Commission found that the tests of teachers' content knowledge--the NTE

Core Battery and the NTE Specialty Area Tests--were not current with the changes in California's reform curriculum. The multiple-choice format of the test questions could not assess the thinking skills in which teachers need to engage their students when thinking about science, mathematics, languages, history, social science and the arts (Wheeler, et. al., 1988). The CTC is currently examining the performance characteristics of the California Basic Educational Skills Test (CBEST). However, because the CBEST was designed to verify a minimum level of proficiency in basic academic skills that should be acquired during elementary and secondary schooling, it is not intended to be a test of teaching ability. The Commission recently initiated several studies to explain why this test continues to be difficult for disproportionate numbers of minority examinees.

The Commission on Teacher Credentialing reappraised the support and evaluation of student teachers, which are now the subject of *Standards of Program Quality and Effectiveness* that the Commission adopted in 1986 and strengthened in 1988. The CTC created teams of teachers and teacher educators to review each teacher education program on the basis of these new standards, which require universities to establish documentary evidence of each teacher's performance in relation to ten uniform criteria of effectiveness.

At the same time, the Commission established panels of other subject-matter experts to develop new examinations of the content knowledge of future teachers. New exams will include subject-matter performance exercises as well as multiple-choice questions, and they will replace the NTE Specialty Area Tests and the NTE Core Battery Test beginning in 1991-92. To ensure that the new exams will be congruent with the *Model Curriculum Standards, K-8 Guidelines, and State Curriculum Frameworks*, the California Department of Education has been an active partner in these changes in subject-matter examinations.

Although these reforms promise to contribute to the effectiveness of the California teacher force in the future, they essentially leave intact the conditions in which beginning teachers work after completing their initial credential requirements. To address the state's induction of new teachers, including the proposals to establish **support systems** and **assessment systems** for first- and second-year teachers, the CTC and CDE are jointly administering the California New Teacher Project, which was authorized by policy legislation and budget appropriations beginning in 1988.

The California New Teacher Project

The California New Teacher Project (CNTP) was created by the legislature in the Teacher Credentialing Law of 1988 (Chapter 1355 of the Statutes of 1988). Charged with exploring innovative methods of new teacher support and assessment, the CNTP has three components: support, evaluation, and assessment. A brief overview of each component and the overall goals of the CNTP are found in this section; the assessment component is described in more detail in the following section.

The support component of the CNTP consists of local pilot projects representing diverse teaching contexts as well as a variety of approaches to supporting new teachers. During the first year (1988-89), fifteen projects funded through a combination of state and local sources participated. The number of projects was increased in the second year (1989-90) to include additional projects either expanding the representation of approaches to new teacher support or continuing programs of district-funded support and receiving funds only to participate in CNTP meetings and data collection efforts. Although these projects are not the only new teacher support programs in California, teachers and administrators in these projects are a key component of the research on alternative methods of new teacher support sponsored by the CNTP.

The evaluation component of the CNTP is designed to investigate the effects of the various methods of support on new teacher effectiveness and retention, as well as cost-effectiveness. The variety of approaches to new teacher support combined with the evaluation of these approaches should help to identify the forms and intensity of assistance that are most effective with new teachers entering the profession. The CTC and SDE have contracted with the Southwest Regional Laboratory (SWRL) to conduct all activities in the evaluation component. The evaluation results of the first two years of the CNTP can be found in two reports: *1988-89 Evaluation Report* (SWRL, 1990) and *1989-90 Evaluation Report* (SWRL, 1991).

Assessment Component of the California New Teacher Project

Many of the reform advocates have criticized the exclusive use of multiple-choice tests in traditional teacher licensure systems. According to many teachers, teacher educators and researchers, multiple-choice questions cannot assess many of the important skills and abilities that characterize proficient, effective teachers. These advocates have recommended that states examine the efficacy of other methods for assessing the capabilities of credential candidates, methods such as on-site observations, oral interviews, structured exercises in assessment centers, and the use of videotaped scenarios and other "prompt materials" in performance assessments. Each of these recommendations was intended to make the assessment of teaching more authentic in relation to teachers' actual duties and requirements. When education policymakers in California faced the choice of assessment methods, however, they quickly discovered that few, if any, of the recommended methods had been pilot-tested or evaluated in practice. The literature on education reform was "long" on suggestions but "short" on evidence of the cost-effectiveness of varied methods of assessing teacher competence and performance.

To help education policymakers with their choice of assessment methods, the assessment component of the CNTP was designed to develop and pilot test innovative forms of new teacher assessment. The evaluation of diverse approaches to teacher assessment is intended to identify the most promising ways in which a comprehensive assessment of teacher candidates could inform the credentialing process and contribute to the quality of teaching. This document reports the analysis of the pilot tests of assessments that were completed during 1990, the second year of the CNTP. The analysis of the first year of pilot testing appears in a previous report, *Assessment Component of the California New Teacher Project: Year One Report*. The pilot tests were administered and analyzed by Far West Laboratory for Educational Research and Development (FWL). The design and purpose of the second year of pilot testing are described in Chapter 2. The 1990 pilot tests differ from the 1989 pilot tests in that the 1990 assessments were specifically commissioned by the California New Teacher Project to increase the diversity of assessment approaches represented in the research and to better reflect California's curriculum and diversity of students.

The Bergeson Act (S.B. 148) which created the CNTP specifically requires that each alternative method of support and assessment be evaluated along the following dimensions:

- effectiveness at retaining capable beginning teachers in the profession;
- effectiveness at improving the pedagogical content knowledge and skills of the beginning teachers who are retained;
- effectiveness at improving the ability of beginning teachers to teach students who are ethnically, culturally, economically, academically, and linguistically diverse;
- effectiveness at identifying beginning teachers who need additional assistance and, if that additional assistance fails, who should be removed from the educational profession;
- the relative costs of the method in relation to its beneficial effects; and

- the extent to which an alternative method of supporting or assessing beginning teachers would, if it were added to the other state requirements for teaching credentials, make careers in education more or less appealing to prospective teachers.

Although both the support and assessment components are guided by relevant state curriculum frameworks and expectations for the pedagogical competence of new teachers, the SDE and CTC have not generated a list of competencies to serve as a common focus for all components of the CNTP. Instead, to increase the variety of methods being evaluated, the assessment component is conducted independently of the evaluation and support components. For this reason, the competencies being measured by the assessment instruments piloted may or may not coincide with the areas of support offered to the new teachers by the support projects. The integration of the lessons learned from the evaluation and assessment components will facilitate an analysis of the relationships and interactions among teacher preparation, support, assessment, and credentialing to suggest whether and how a program of support and assessment for new teachers should be developed.

In examining current approaches to teacher assessment, CTC and SDE staff found few assessment approaches that are closely related to the tasks that teachers perform in the course of their work. This lack has led nationally to the development of alternatives to multiple-choice tests, which historically have been the dominant form of large-scale teacher assessments. The alternatives are often referred to as "innovative" or "performance-based" assessments because of their emphasis on direct measurement of actual teacher performance.

A variety of performance-based teacher assessments has been developed in recent years, including a number of observation instruments which have been adopted as teacher credentialing requirements in other states. However, many of these instruments are very prescriptive in terms of teaching style. Since California classrooms are extremely diverse, instruments which tend to promote only one or a few teaching styles are inappropriate for use in assessing California teachers. For this reason, the CNTP is designed to evaluate the degree to which various assessment approaches measure the ability of teacher candidates to teach a wide variety of students.

The Bergeson Act reflects an emerging design for California's assessment of teacher candidates in four areas: (1) basic academic skills; (2) subject matter knowledge; (3) subject specific pedagogy; and (4) general pedagogy. The CBEST has been judged to be suitable for assessing candidate performance in the first area (Watkins, 1985), and revisions are under way in the second area in tests that measure subject matter knowledge of elementary teachers (NTE Core Battery) and secondary teachers (NTE Specialty Area Tests). The third and fourth areas, which were judged to be most effectively assessed after candidates have had some experience in conducting their own classrooms (i.e., in the first year or two of teaching), are the primary focus of the CNTP. The CNTP aims to identify promising, cost-effective assessments of subject-specific pedagogy and general pedagogy, especially in the following areas: Secondary English, Secondary Mathematics, Secondary Life Science, Secondary Physical Science, Secondary Social Science, and Elementary Teaching.

Because of the high interest in teacher assessment among educators in recent years, together with a growing recognition of the limitations of the multiple-choice approach, new assessment approaches are being developed, and old approaches are being revised. New approaches include the use of videotapes, written vignettes, structured interviews, structured simulations, and reviews of portfolios of a teacher's work. More traditional approaches such as classroom observation are being revised and refined so as to go beyond the checklist format and to move toward an instrument which provides rich information with strong diagnostic potential.

In planning the research to be conducted in the assessment component of the CNTP, staff from the CTC and SDE considered both the high cost of assessment development and the desirability of evaluating a wide variety of assessment approaches. Many "innovative" assessment instruments are in the initial stages of development, and could only serve as initial prototypes for exploring the potential of an assessment approach, rather than as state-of-the-art instruments reflecting a long period of experimentation within that approach. The most promising state-of-the-art instruments representing assessment approaches in later stages of development were, for the most part, pilot tested during the first year of the CNTP. Therefore, to maximize the information to be gathered while minimizing developmental costs, the assessment instruments commissioned for pilot testing the second year were not required to be fully developed products with well established validity and reliability. Instead, the second year's pilot testing was designed to yield information about the strengths and weaknesses of assessment **approaches** for which the specific instruments serve as exemplars. The purpose of the pilot testing is not to consider particular instruments for adoption, but to identify promising approaches to the assessment of teachers, to guide future selection and/or development of instruments which are tailored to the California context. Consistent with this purpose, assessment prototypes were piloted on a small scale with a thorough trouble-shooting process in order to learn as much as possible about the strengths and weaknesses of each approach before incurring the expense of large-scale field tests.

1990 Pilot Testing

The instruments pilot tested during 1990 were commissioned during the first year of the CNTP to represent subject matter areas or assessment approaches which had been insufficiently explored. These instruments and the approaches which they represent are as follows:

Instrument	Approach
Structured Simulation Tasks for Secondary Life/General Science Teachers	Structured Simulation Tasks
Laboratory Science Assessment	Subject-Matter Specific Classroom Observation
Language Arts Pedagogical Knowledge Assessment	Videotaped Teaching Episodes
Structured Simulation Tasks for Secondary English Teachers	Structured Simulation Tasks
Secondary English Assessment: Assessment Center Activities	Performance-Based Exercises
Secondary English Assessment: Portfolio Activity	Classroom Portfolio
Semi-Structured Interview in Secondary Social Studies	Semi-Structured Interview
Assessment of Competence in Monitoring Student	Structured Simulation Tasks

The evaluation of the various components (e.g., logistical requirements, prompt materials, scoring criteria, training of assessors and/or scorers) of the instruments was intended to provide information about the strengths and limitations of the assessment approaches which the specific instruments represented. The pilot tests were not expected to yield definitive measurements of the psychometric properties of the instruments because the prototypes had not been sufficiently developed for that to occur. This focus on troubleshooting allows small-scale pilot testing, requires fewer resources, and considerably increases the number of assessment approaches which can be examined. The goal of the pilot tests is to suggest whether or not it is advisable to invest additional resources in the development of assessments resembling those piloted.

This document is the final report and analysis of the administration and scoring of the assessment instruments pilot tested in 1990. The next chapter describes the pilot test design and the processes used to evaluate the assessment approaches which were examined in 1990. In the chapters that follow, each of the assessment instruments is described, with each chapter including a discussion of the ease of administration, scoring, content and format, costs, and technical qualities of the instrument. The instruments are presented in the following order: Structured Simulation Tasks for Secondary Life/General Science Teachers, Laboratory Science Assessment, Language Arts Pedagogical Knowledge Assessment, Structured Simulation Tasks for Secondary English Teachers, Secondary English Assessment: Assessment Center Activities, Secondary English Assessment: Portfolio Activity, Semi-Structured Interview in Secondary Social Studies, and Assessment of Competence in Monitoring Student Achievement in the Classroom. The report concludes with a summary of strengths and weaknesses of the assessment approaches represented by these instruments, conclusions about the effective design of training for assessors and/or scorers, and an identification or augmentation of policy issues beyond those discussed in the first year report which will affect the design of a teacher assessment system.

CHAPTER 2:

PILOT TEST DESIGN AND ANALYSIS

This chapter describes the design and analysis of the pilot tests of prototypes representing various assessment approaches. Different sections describe the source of instrumentation, the sampling plans, sources of information for evaluating the instruments and the assessment approaches, methods of data reduction and major categories of analysis. Deviations from the design due to unanticipated events will be described in following chapters which focus on the individual instruments.

Design of Pilot Tests

This section on the design of the pilot tests describes the sources of instrumentation and the sampling plans. Procedures for data collection and analysis will be described in the sections on data collection and data reduction.

Sources of Instrumentation

In the first year of the project, the prototype instruments that were pilot tested were selected on the basis of their representation of state-of-the-art development of innovative assessment approaches. For this second year of pilot testing, the Interagency Task Force commissioned the development of additional prototypes through a competitive bidding process. It was intended that these new prototypes would be more congruent with the California Model Curriculum Guides than those pilot tested in the first year of the project which were developed for other states or a national audience. The new prototypes commissioned were also chosen to represent a variety of assessment approaches. Each will be described separately.

The Structured Simulation Tasks for Secondary Life/General Science Teachers is a set of structured simulation tasks to which teachers respond in writing. The tasks are chosen to represent important responsibilities which differentiate more and less competent beginning teachers. This assessment was developed by the RAND Corporation in Santa Monica, using the same process that was used to develop performance tasks for the Bar examination to license lawyers. The current set of tasks does not represent a complete assessment, but rather prototype tasks that may eventually be incorporated into a complete assessment.

The Science Laboratory Assessment, developed by the RMC Corporation in Mountain View, California, combines classroom observation with structured interviews to measure both general pedagogical skills and instructional skills in a science laboratory setting.

The Language Arts Pedagogical Knowledge Assessment (LAPKA) was developed by the Northwest Regional Educational Laboratory (NWREL) in Portland, Oregon. It uses videotapes of teachers instructing small groups of their students to portray a variety of approaches to language arts instruction. The videotapes are stopped at various points to pose questions to which the teachers respond in writing.

The Structured Simulation Tasks for Secondary English Teachers is a set of structured simulation tasks to which teachers respond in writing. The tasks were designed to elicit demonstrations of knowledge specific to secondary English teachers. The set of tasks was developed by the RAND Corporation in much the same way as the set of structured simulation tasks for science teachers was developed (see above).

The Secondary English Assessment: Assessment Center Activities, developed by San Francisco State University, is comprised of three performance-based activities, each of which requires the teacher to demonstrate (or "perform") a different skill or ability. The activities were originally developed as one part of a two-part assessment (the other part being a portfolio activity), but because the two parts were administered at different times of the year, each part was analyzed as a separate assessment.

The Secondary English Assessment: Portfolio Activity was also developed by San Francisco State University. For this activity, a teacher's skills are assessed through a portfolio format. Specifically, the teacher is given three months to plan and conduct a three- to six-week teaching unit and to compile a portfolio that documents the activities of the unit.

The Semi-Structured Interview in Secondary Social Studies (SSI-SSS) is a performance assessment developed by the Connecticut State Department of Education. The assessment targets a beginning teacher's knowledge in the subject area of social studies, exploring a teacher's thought processes as he or she makes instructional decisions for students. For the assessment, the teacher completes four tasks, each of which is followed by a semi-structured interview designed to elicit the teacher's rationale for the choice(s) made.

The Assessment of Competence in Monitoring Student Achievement in the Classroom is a set of structured simulation tasks which aims to assess an elementary teacher's ability to measure student achievement. The teacher responds in writing to each task. It was developed by a second team from NWREL in Portland, Oregon, based on a decade of research and development of training on the topic of classroom assessment.

The developers of the assessments provided guidelines for administration and, except for the various activities of the Secondary English Assessment, supervised the training of scorers and/or observers. Experienced English educators who had participated as scorers in an earlier administration (i.e., one conducted by the developer) of the Secondary English Assessment served as trainers of administrators/scorers for the assessment center activities and the portfolio activity. All observers and scorers of the assessments were recruited by FWL staff; some of these had previously participated in the development of the assessments.

Sampling Plans

Our goal was to obtain a broad sample of teachers representing both genders as well as a variety of ethnicities and teaching contexts. In addition, we desired participation from teachers representing the range of grade levels included in the credential that was the focus of the assessment.

In the case of the Assessment of Competence in Monitoring Student Achievement in the Classroom, two districts were recruited to provide groups of elementary teachers to participate in the assessment. For the other assessments, recruitment of individual teachers was necessary. We began the sample selection process by assembling lists of possible participants within each project in the California New Teacher Project (CNTP). For the science assessments, it was apparent that teachers outside the CNTP would need to be contacted in order to reach the desired sample size. In these cases, personnel offices of virtually every school district in the greater Bay Area and Los Angeles areas were contacted to obtain lists of teachers. Once these lists were completed, the characteristics of grade level, school context (e.g., inner city, suburban, rural), gender and ethnicity were considered in selecting teachers to contact for possible participation in an assessment.

Other factors influenced the choice of teachers for various assessments. For example, for the Structured Simulation Tasks for Secondary Life/General Science Teachers, the Structured Simulation Tasks for Secondary English Teachers, and the Language Arts Pedagogical Knowledge Assessment, it was important that teachers be located reasonably close to a central assessment site. On the other hand, for the Science Laboratory

Assessment in which teachers were individually observed, teacher selection entailed balancing the goals of minimizing observer travel costs, matching teacher and observer availability, and obtaining a broad sample. Still another factor was involved in the selection of teachers for the Secondary English Assessment: Assessment Center Activities, which was administered at a single site during one week in the summer. Since statewide representation in all assessments was desired, the budget provided for half the teachers to travel to the assessment by air and half by local transportation. This budgetary constraint guided sample selection for the assessment.

Although we wanted to maximize variation in the characteristics of teachers selected, our ability to do so was limited by the information which we had about project teachers, the time required to recruit nonproject teachers, and the small samples. Information on the ethnicity of teachers was available for many of the projects, but there were few nonwhite teachers, precluding the selection of a significantly large subsample. Our information on school context was limited to our knowledge of the districts participating in the various projects, augmented by conversations with the Project Directors and teachers.

Even though obtaining a broad sample of teachers was a goal, this was not possible for all assessments. Considerations of administration costs and geographic dispersion of teachers led to an underrepresentation of rural teachers in most assessments. The recruitment of minority teachers was a priority, but locating minority teachers proved to be difficult. The number of minority teachers participating in the assessments ranged from three to fifteen. The characteristics of teachers in the samples are described in more detail in the chapters that focus on specific instruments.

This section describes our procedures for data collection and reduction, as well as the key analytic categories focusing on specific aspects of instruments. The data collected also served as a basis for judging the potential of the assessment approach which the particular instrument utilized.

Data Collection

Since the same means of data collection were used for all assessment instruments, they will be discussed together. Several sources of data were used:

- evaluation feedback forms completed by teachers who participated in the pilot tests;

- evaluation feedback forms completed by the observers and scorers;

- observations of the administration of each assessment and the training of observers and scorers recorded in field notes by FWL staff;

- scores that reflected the performances of participating teachers on the assessment instruments;

- review of instruments or portions of instruments by an expert on teaching diverse students; and

- the most recent relevant Curriculum Guide(s) and/or Framework(s) and the California Standards for Beginning Teachers.

Using the list of analytic categories and the evaluation feedback forms developed during the first year of the project, FWL staff developed separate evaluation forms for each group of participants (e.g., teachers, scorers) which were tailored to specific assessment instruments. These forms were given to teachers upon the completion of each assessment, except in the case of the classroom observation instrument, where they were mailed. Observers and scorers also returned completed forms with their invoices for payment. Since the emphasis in the pilot tests was on trouble shooting, the evaluation feedback forms focused on critical evaluations of the instruments with respect to the analytic categories described in the next section. Most of the questions required yes/no or fixed response answers with spaces provided to elaborate.

Field notes were taken during observations of the assessment administrations. FWL staff conducted most administrations of the assessment instruments, and accompanied one observer during the use of the observation instrument. FWL staff also observed the training of observers and scorers. For the Assessment of Competence in Monitoring Student Achievement in the Classroom, FWL staff also served as participant observers for scoring to obtain a more complete understanding of the performance of the assessment instruments.

The content of each prototype was compared to all of the relevant California Model Curriculum Guides and Frameworks, and with the California Standards for Beginning Teachers. The Model Curriculum Guides and Frameworks are recent documents produced by subject matter panels convened by the California State Department of Education. Reflecting a consensus among panel members on the content and philosophy of instruction, these documents are expected to guide curriculum development and instruction in the subject in California public schools. If there were two or more Guides or Frameworks addressing a particular subject area, the most recent one available was used.

The California Beginning Teacher Standards are standards that define the level of pedagogical competence and performance that the Commission on Teacher Credentialing expects the graduates of credential programs to attain as a condition for program approval. These standards--Standards 22 through 32--are listed in *Standards of Program Quality and Effectiveness, Factors to Consider and Preconditions in the Evaluation of Professional Teacher Preparation Programs for Multiple and Single Subject Credentials.* (Other standards address more general program requirements; Standards 22 through 32 focus specifically on candidate competencies.) Although these are standards for teacher preparation programs and not teacher candidates, they identify the knowledge and skills that beginning California teachers are expected to attain.

Data Reduction

Data reduction techniques varied with the data collection method. Fixed-response questions on the evaluation feedback forms completed by all participants in the pilot tests (e.g., teachers, observers, scorers) were tabulated. Open-ended responses and elaborations were compiled. Responses which either stated a common viewpoint well, or which provided an additional perspective, were highlighted for possible quotation in the reports. For the fixed-response questions where elaboration was invited, the focus was on identifying weaknesses in the instruments and on soliciting suggestions for improvement. Therefore, teachers were only asked to comment on negative responses, so there were many more negative evaluations available for quotation than positive ones.

Field notes were reviewed for relevant information that addressed the analytic categories, and these notes were incorporated into the chapters about specific instruments.

When numbers were large enough to permit analysis of scores by subgroup, the following comparisons were made: male/female, minority/nonminority, and teachers at different grade levels and in different locations (urban, rural, inner city).

At least a portion of each assessment was scored by two people to assess inter-rater reliabilities. Scores were also used to estimate the internal consistency of an instrument.

The Model Curriculum Guides and Frameworks were examined by FWL staff. Their professional judgments were used to draw conclusions about each assessment instrument's extent of coverage and congruence with the relevant Guide or Framework. The reasoning underlying these judgments is described in detail in the chapters on the specific prototypes.

Overview of Analytic Categories

The same general analytic categories were used to appraise all assessment instruments. They included: administration, content, format, cost analysis, and technical quality. These categories and their subcategories are discussed below.

Administration of assessment. This category included consideration of the logistics, security needs, and training of observers and scorers for the particular assessment instrument. Generally, this category generated information required to estimate administrative requirements and cost projections. The logistics required for administration predict the ease of administration if the assessment approach were to be implemented on a statewide basis. The more complicated the logistical requirements, the more expensive the assessment is to administer. Security needs impact not only logistical requirements, but also the frequency with which the instrument must be revised for statewide administration. Consideration of the training of observers and scorers suggests the degree of difficulty to be anticipated in recruiting people with the required professional expertise, and the time required to prepare personnel to administer and score the particular assessment instrument.

Assessment content. This category addressed the specific instrument's congruence with the relevant Curriculum Guide or Framework, and the extent to which the California Standards for Beginning Teachers were covered. It also included an examination of the content of the assessments along the following dimensions: job-relatedness, appropriateness for beginning teachers, appropriateness across varying teaching contexts, fairness across different groups of teachers, and general appropriateness of the assessment approach represented by the prototype as a method of assessing teachers. Comparison of the assessment content with the relevant Curriculum Guide and the California Standards for Beginning Teachers was necessary to determine whether the assessment approach was compatible with the instructional philosophy underlying the various California curricula and the competencies specified for teacher candidates. Since one common criticism of teacher assessment instruments is that scores have not been shown to be closely related to specific teaching competencies, job relevance was included as an analytic category. The more closely the assessment tasks resemble the activities that teachers do in the course of their teaching duties, the higher the potential relationship of scores to actual teaching competencies.

Since the CNTP focuses on the assessment of teachers early in their teaching career, it is important to judge the appropriateness of each assessment in terms of performance expectations and perceived difficulty for teachers at this stage of career development. Appropriateness across contexts is particularly important for California, since it has a wide diversity in student populations. The issue of fairness across groups of teachers relates to the potential for bias with regard to any particular group of teachers (e.g., gender, ethnicity).

Assessment format. This category included the general clarity of orientation materials, directions for completing the assessment, and scoring criteria. In order for the performance of candidates to reflect their true competencies, it is essential that all candidates have clear and accurate expectations of the performance that is expected of them. This is not possible when teachers are uncertain as to what they are being asked to do. It is equally important that scorers have a clear understanding of the criteria by which they are judging a teacher's performance.

Cost analysis. Based on the pilot testing experience, we attempted to project the costs of a statewide administration and scoring of an instrument which resembled the prototype tested. We also have reported the costs for the developers to develop these prototypes and for FWL's pilot testing. The developmental costs experienced to date provide a rough basis for judging the developmental costs for assessment approaches like these.

Technical quality. This category discussed the work performed to date in the development of the prototype, together with data estimating the reliability and validity of the instrument.

This chapter has outlined the general design for the 1990 pilot tests in the assessment portion of the California New Teacher Project. The following eight chapters discuss each of the assessment approaches pilot tested as represented by the various instruments: structured simulation tasks (the Structured Simulations Tasks for Secondary Life/General Science Teachers, the Structured Simulation Tasks for Secondary English teachers, and the Assessment of Competence in Monitoring Student Achievement in the Classroom), classroom observation (the Science Laboratory Assessment), videotaped teaching episodes that require written responses to questions (the Language Arts Pedagogical Knowledge Assessment), a semi-structured interview (the Semi-Structured Interview in Secondary Social Studies), performance-based assessment center exercises (the Secondary English Assessment: Assessment Center Activities), and a portfolio (the

Secondary English Assessment: Portfolio Activity). Each of these assessment approaches is further described and analyzed in the final chapter of this report.

CHAPTER 3:

STRUCTURED SIMULATION TASKS FOR SECONDARY LIFE/GENERAL SCIENCE TEACHERS

The Structured Simulation Tasks for Secondary Life/General Science Teachers, developed by the Rand Corporation, is a set of structured simulation problems to which a teacher responds in writing. A complete assessment was not developed; development work focussed on the construction of prototypic tasks, each of which may eventually be combined with other tasks to form a complete assessment. (To simplify references to these prototypic tasks, however, they will be referred to collectively as "the assessment.") For this pilot test, all stimulus materials were in written form, although the assessment developer sees videotape as a possible alternative stimulus.

To facilitate future development of parallel tasks, the construction of each task begins with the design of a blueprint for production, which the developer terms a "shell." No two shells have exactly the same features and components. However, most shells provide the following:

- a general description of the activity or types of activities that will be present in a task, (e.g., "grade a set of student papers that exhibit at least five of the following characteristics...") and the general directions to candidates;

- things that can be built into a task that candidates should attend to in specified ways (and which can be scored with respect to whether the candidate did or not attend to them, e.g., one answer is symptomatic of a common learning difficulty or disorder);

- the types of materials candidates will receive (both in advance of the test and at the test site); and

- any special features of the context that need to be explained.

Many different items or case situations can be generated from the same shell.

For this pilot test, the following five tasks were fully developed, from stimulus materials through scoring criteria:

(1) **Applying Effective Instructional Techniques.** A teacher reads a simulated transcript containing several lesson segments from a single class, and identifies appropriate and inappropriate actions and statements made by the teacher in the script, commenting on why each is appropriate or inappropriate.

(2) **Teacher as Curriculum Decision-Maker.** This task consists of two parts. In the first part, the teacher combines a subset of given activities (e.g., lectures, laboratories, films, tests) into a two-week (ten day) lesson plan to achieve a given set of student objectives for the classroom described. In the second part, the teacher also provides a rationale for the overall plan.

(3) **Parent/Student Letter.** This task also consists of two parts which relate to drafting a letter regarding a science course, for which descriptions of the course and the students are provided. The letter is to be sent to parents and students at the beginning of the school year. In the first part, the teacher lists reasons why the course would be important and of value to students. In the second part, the teacher develops an outline of additional topics, including any required by law, to be included in the letter.

(4) **Lesson Planning.** The two parts of this task focus on a specific lesson in a unit. A description of students, the instructional goals of the unit, and the other lesson topics in order of presentation are provided. In the first part, the teacher analyzes the strengths and weaknesses of three alternative lessons designed to fill the missing slot in the sequence of lessons in the unit. In the second part, the teacher designs a more effective lesson and describes its strengths and weaknesses. The teacher is free to modify one of the lessons provided or to design a new one.

(5) **Classroom and Facility Safety.** This task consists of three parts. In the first part, the teacher provides a list of categories of activities (excluding facilities) that teachers can do alone or with their students at the beginning of the year to promote classroom laboratory safety. In the second part, the teacher lists specific actions that would promote safety and prevent or reduce the likelihood of accidents in a specific laboratory activity with the classes of students

described. In the final part, the teacher identifies safety hazards in a drawing of part of the classroom.

In addition, six other tasks were partially developed, ranging from the shell stage to a revised draft of the shell, stimulus materials, and scoring criteria. These other tasks addressed the following topics: the transition to the laboratory, common scientific misconceptions, understanding student behavior, using computers as tools, evaluating student performance, and meeting special needs. However, only the five tasks described above were pilot tested.

Two forms of **Applying Effective Instructional Techniques** were pilot tested. Each form contained four lesson segments, and six lesson segments were developed. Two segments were common to both forms, and each form contained two of the remaining four segments.

Administration of Assessment

The administration of the assessment, the assessment content, and the assessment format are discussed below. The discussion of the Secondary Life/General Science Assessment concludes with a summary of our evaluations of its potential as a prototype for further assessment development.

Overview

The Secondary Life/General Science Assessment was administered at five sites in the Bay Area and the greater Los Angeles area between June 2 and June 23, 1990. As seen in Table 3.1, a total of 65 teachers participated, the majority of whom were female. The teachers included sixteen minority teachers. A little over half of the teachers taught in either a middle school or a junior high school; two additional teachers had teaching assignments split between junior high and high schools. Approximately three-quarters of the teachers graduated from traditional teacher preparation programs. Nearly all the remaining teachers participated in intern programs, where they received their pedagogical training while assuming sole responsibility for their classes of students. The two teachers whose training fell into the "other" category received teacher training through the Peace Corps.

TABLE 3.1

PILOT TEST PARTICIPANTS

SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

(Number of Teachers = 65)

Descriptive Characteristics of Participants	Distributions of Participants	
	Form A (N=32)	Form B (N=33)
Gender		
Male	11	14
Female	21	19
Ethnicity		
Asian	2	4
Black	1	1
Hispanic	5	0
Native American	0	2
White	24	24
Other	0	1
No Response	0	1
Grade Level		
Middle/Junior High School	18	19
High School	12	14
Both Junior and High School	2	0
Source of Teacher Preparation		
Intern Program	7	7
Regular Credential Program	24	24
Other	0	2
No Response	1	0

With the exception of the intern teachers, participating teachers were their first or second year of teaching. The intern teachers were either in their second and final year of training or in their first year of teaching following completion of the program.

Two different forms of the assessment, with two segments of one task in common, were piloted. Thirty-two teachers completed **Applying Effective Instructional Techniques** (Form A), **Teacher as Curriculum Decision-Maker**, and **Parent/Student Letter**. Thirty-three teachers completed **Applying Effective Instructional Techniques** (Form B), **Lesson Planning**, and **Classroom and Facility Safety**. Teachers completed the tasks in approximately half a day: One hour was allotted for the first task, ninety minutes for the second, and forty-five minutes for the third.

Logistics

Administration required the following logistical activities: identifying a sample of teachers, sending orientation materials to teachers, administering the assessment, and acquiring evaluation feedback from the teachers.

Identifying teacher samples. The California New Teacher Project contained too few science teachers to provide a sample for the assessment. Therefore, unlike the other pilot tests, most participants in this assessment were Non-project teachers. We focussed on the state's two largest urban areas, the greater Los Angeles area and the Bay Area, to locate a sufficient number of first- and second-year life science teachers. The personnel office of most school districts in these two geographic areas was contacted and asked to either supply the names and school sites of any appropriate teachers or, if their policy prevented the release of names, to forward a letter to the appropriate teachers inviting them to participate in a pilot test. When the names of teachers were obtained, a letter was sent to them followed by a telephone message left at their school site inviting them to call collect for more information. Many of the teachers identified turned out to be in their first or second year in the district, but had more than two years experience. The majority of the bona fide first and second-year teachers contacted agreed to participate in the assessment. More teachers than needed were scheduled to participate to allow for some attrition.

Sending orientation materials. The assessment developer provided the orientation material for the teachers, which consisted of brief descriptions of six possible tasks which they would be asked to do (including one which was not pilot tested) and a list of 27 possible science topics which might serve as the focus for the tasks. In addition, teachers received a letter briefly describing the California New Teacher Project and its Assessment

Component, and directions to the assessment site. Teachers were paid $ 80 for participating in the assessment and completing an evaluation form.

The assessment was designed to be administered to large groups by a test administrator who distributed and collected materials, announced the start and end of each task, and monitored the teachers to prevent cheating. No special training or background in science was needed, as the instructions were designed to be self-evident.

The only requirement which differed from those of traditional group-administered tests was that of sufficient surface area (e.g., individual desks or a number of tables) to spread out a number of materials. Facilities which fit this requirement proved to be easy to locate, and included classrooms used by a district for professional development, a large conference room, and a room in a medical center set up for classroom instruction.

Each assessment began with a ten-to fifteen-minute overview of the research design underlying the California New Teacher Project. Teachers were given the option of a five to fifteen minute break between tasks, but usually opted to limit the break to five minutes to finish earlier. In the overview, which was similar for all pilot tests (except the classroom observation assessment), the following topics were covered: (1) the purpose of the pilot testing and descriptions of the spring pilot test activities; (2) identification of the assessment developer and distinctions between the roles of the assessment developer and FWL; (3) the confidentiality and use of the results; and (4) a description of the evaluation form which teachers would complete at the end of the assessment.

Test materials were distributed in three manila envelopes, with each envelope containing a single task. The envelopes were labeled with both the task code and an ID number. Teachers were instructed to record that ID number on the test materials and the evaluation form.

Security

It is the position of the test developer that once the test is given, its security is compromised, and new forms of the tasks must be developed. Therefore, security precautions coupled with the fiscal need to reduce development costs dictate that it be administered to large groups in various locations on the same date. Facilitating the development of parallel tasks to maintain both security and fairness led the developer to conceive of the "shell" system for generating tasks.

The tasks would almost certainly be memorable. Some are more amenable than others to coaching through memorization of acceptable answers, e.g., Part II of the **Parent/Student Letter**, where the teacher lists topics other than course content to be covered in the letter. This list would be similar regardless of the course content described. However, learning test-taking techniques and common answers would be less useful for a task such as **Teacher as Curriculum Decision-Maker** or parts of **Classroom and Facility Safety**, where answers depend on the content, and the teacher needs to be able to apply general principles in light of the specific content portrayed.

Assessors and Their Training

Two members of the FWL staff administered the assessment. No training was provided other than instructions about times for the tasks and suggested breaks. Before the first assessment, the two staff members designed the assessment schedule, including time for the overview and the evaluation form. No need for further training in test administration was detected by the staff, who were experienced in conducting assessments, although standardized guidelines for dealing with possible situations, e.g., a test-taker becoming ill during the test, would be needed for statewide administration.

Teacher and Assessor Impressions of Administration

Teachers responded favorably when asked their impressions of the arrangements for administration, including scheduling, room arrangements, and distance to travel to the assessment site. Fifty-seven of the sixty-five teachers (88%) responded that the arrangements were reasonable. Comments critical of the arrangements addressed travel distance, early morning traffic coinciding with the time at which the test was scheduled, and security at one site where a teacher discovered vandalism to his car which was assumed to have occurred during the testing.

Scoring

The discussion of scoring addresses the scoring process, the scorers and their training.

Scoring Process

The scoring guide is built into the tasks during development. For example, in **Lesson Planning**, scorers do not grade a candidate's ability to distinguish between appropriate and inappropriate plans. Neither are teachers asked to list the rules for good plans. Instead, teachers evaluate actual plans, and scorers determine whether the teacher responded appropriately to a specific situation in which the ability to evaluate plans was needed. The scoring process differs slightly among tasks, but is generally based on correct identification of appropriate or inappropriate items built into the stimulus materials. Points are deducted for responses which are clearly wrong. For several teachers, this resulted in a negative score on one or more parts of a task. Scoring guides are modified after the task is administered, such as when the examinees see certain strengths or weaknesses that were not anticipated by task developers. The scoring system for each task will be discussed separately in more detail.

The task **Applying Effective Instructional Techniques** consisted of four simulated segments of a single class. The segment included both a transcript of teacher/student conversations and, when needed to interpret the transcript, a description of what the teacher or the students were doing. Teachers were asked to identify both appropriate and inappropriate actions by the teacher and to briefly comment on them.

Certain appropriate or inappropriate actions were built into the script when it was constructed, such as building upon previous instruction or reprimanding one student and not another for similar behavior. Scorers were presented with a list of these appropriate and inappropriate actions built into the script. A few additions to the list based on teacher responses were made during the initial training to score the task. In the case of any teacher responses not already covered by the list, scorers were instructed to base their judgement on the previously identified examples. Teachers received one point each for every appropriate or inappropriate action they correctly identified. If a teacher identified an action incorrectly, i.e., said it was appropriate when it clearly was not or vice versa, one point was deducted. Some teacher comments were labeled "neutral" during the training, as when the comment was judged to be too vague or when the teacher went beyond the script in making assumptions about the teacher behavior. These "neutral" comments received a "zero" score. However, when teacher assumptions clearly contradicted the information provided in the script and accompanying materials, a point was deducted.

The two parts of **Teacher as Curriculum Decision-Maker** were scored differently. The first part consisted of choosing a subset of activities provided and arranging them into a two week unit of instruction given specified unit objectives and a description of the group of students in the classroom. Activities were divided into the following categories: Lecture/Discussion, Demonstration, Reading, Laboratory Activity, Film/Video, Student Worksheet, Homework-in-Class, and Testing/Evaluation. This part was scored using an algorithm which awarded a teacher 100 points and then deducted varying points for the following: (1) incorrectly sequencing antecedent and subsequent activities when both were included in the unit; (2) omitting necessary prerequisites for activities included; (3) failing to cover one or more unit objectives (which were provided); (4) including topics which were tangential to the unit; (5) including activities which were too difficult for the class described; (6) failing to assemble enough activities, including homework in class, to cover a 50-60 minute class period (times were provided for each activity except homework-in-class); (7) failing to include a variety of activities over the entire unit (i.e., not having lessons predominantly composed of activities from a single category); (8) failing to have a variety of activities each day; (9) using too much instructional time for homework in class; (10) assigning too much weekly homework; and (11) giving too many tests within the two-week period. When a large number of points were available for deduction for any of the above, a ceiling on the number of points deducted was established. For example, although 35 points were possible for deduction for incorrectly sequencing activities, a maximum of 10 were deducted. This rule had to be applied in several cases.

In the second part of the **Teacher as Curriculum Decision-Maker** task, the teacher provided a rationale for the activities chosen for the unit. A set of possible rationales was devised prior to the scoring training by the task developers; additions were identified during the training. A teacher received from 0 to 3 points per rationale mentioned, depending on whether it was appropriate and, if appropriate, on the depth with which it was explained. Clearly inappropriate rationales received a deduction of one point. The **Parent/Student Letter** task was scored similarly, with the first part covering reasons for taking the course scored with 0 - 2 points per reason, and the second listing additional topics to be covered in the letter scored with 0 - 3 points per topic.

The **Lesson Planning** task had two parts. In the first part, the teacher listed strengths and weaknesses of three alternative lessons which filled a gap in a specific unit of lessons and, together with the other lessons, addressed a set of unit objectives. The unit plan, except for the missing lesson, and the unit objectives were provided. Teachers received one point for each distinct but appropriate strength or weakness; one point was deducted for inappropriate responses. Scorers worked from a previously established list of strengths and weaknesses, but were free to award points if, in their professional judgement, they believed that the candidate response, though not on the list, was valid.

In the second part of **Lesson Planning**, the teacher provided an alternative design for the missing lesson and described its strengths and weaknesses. The description was to include student performance objectives, key concepts to be taught, the sequence of classroom events with anticipated times of completion for each activity, and homework. The teacher responses were scored for both organization and content. Each was scored on a four-point scale, with 0 points for responses judged to be among the worst, 1 for below average responses, 2 for average responses, 3 for above average responses, and 4 for responses judged to be among the best.

The **Classroom and Facility Safety** task had three parts. In the first part, the teacher was asked to list categories of activities that teachers can do with their students at the beginning of the year to promote laboratory safety for the term (excluding facilities). In the second part, the teacher was again asked to list specific things to do to promote safety and reduce the likelihood of accidents, but the context was that of a specific laboratory activity. In the third part, the teacher was given a diagram of a section of a science classroom and asked to identify safety hazards. Parts one through three were scored in a similar manner. Teachers received 1-2 points for each appropriate and distinct response they listed, depending on the specificity and/or depth with which the category was described. Unlike some of the other tasks which asked scorers to use their judgement to award one versus two points, specific criteria were provided to differentiate between the one-point and two-point responses. One point was deducted for each inappropriate response, with a maximum of two points deducted.

This assessment was designed to produce a licensure decision in the most cost-effective yet reliable manner possible. The process of scoring by creating a set of proper responses and measuring how many the teacher identified (and allowing for original ones) captures how well a teacher does or does not do a designated task. If the ultimate set of tasks which constitute the assessment are deemed to represent a sufficiently broad sample of tasks that are critical to success in teaching, the test should be sufficient for purposes of licensure. However, since there is little information on the extent to which a teacher exhibits specific teaching competencies either within or across tasks, this assessment is less useful for yielding diagnostic information for staff development or beginning teacher support purposes.

Scorers and Their Training

Scorers were recruited mainly from the task development team. As a result, three out of the four scorers had participated in the development of the tasks. Their extent of participation ranged from conceptualization and review of materials to major development work on one of the tasks piloted. The scorers included two current science teachers and two district science specialists who worked with beginning teachers.

Scorers were asked the degree of knowledge of science and of science teaching needed to accurately score the assessment. Their general consensus was that minimal knowledge was needed, although two of the four scorers specified a context of an experienced science teacher overseeing a small group of scorers. FWL staff believe that the degree of knowledge of science and science teaching needed for accurate scoring varies from task to task. Both of the two FWL observers of the scoring training were experienced teachers, but were not trained in science. They each found it easier to judge general pedagogical principles (e.g., built on previous instruction) than to judge aspects that were more content-related (e.g., the appropriateness of the homework assignment in **Lesson Planning**). For many tasks, several additions were made to the list of acceptable responses during the scoring of ten sample responses. A non-science teacher might not be able to recognize acceptable responses which were not on the original list. No data are available to estimate the frequency with which novel acceptable responses occurred.

Training for scoring all tasks was similarly structured but conducted separately. To calibrate the scorers (i.e., make sure each was scoring similarly), the following process was used: First, copies of the stimulus materials and scoring guide were distributed, and scorers read through them. The trainer then asked the scorers to score one teacher response. The response was then analyzed, point by point, and scoring of each part was discussed. When scorers disagreed with the trainer, the rationale underlying the scoring was discussed, and a decision was reached on how to score similar responses. Sometimes this involved a greater understanding of how to apply and/or refine the existing scoring criteria; sometimes this entailed adding a response category to the original set for which credit was to be given. This process was repeated until the responses of approximately ten teachers had been scored by the group. (The developer indicated that when he trained similar groups of scorers for statewide assessments, 50 responses were used for the calibration phase.) Scorers then evaluated teacher responses on their own. Each teacher was scored by two scorers. A trainer checked the ratings for each individual teacher. If scorers were two or more points apart in their total score for a task part, then they were asked to confer and resolve the scoring discrepancy within one point.

For one or two subparts of some tasks, (**Lesson Planning**, **Applying Effective Instructional Techniques**, and **Classroom and Facility Safety**), the original scoring criteria were extensively revised due to problems in their implementation. Problems in the stimulus materials were identified in the course of scoring. These problems tended to be a lack of information that allowed teachers to make wrong assumptions or a need for more specificity in directions to more clearly indicate the focus of the desired response. Generally, these

BEST COPY AVAILABLE

problems were minor. The stimulus materials needing major revision were one part of **Applying Effective Instructional Techniques** that focused on teaching students of diverse cultures. The problems identified in the stimulus materials and scoring criteria suggest a need for more extensive pilot testing prior to actual administration.

All scorers evaluated their training as "very good," the highest rating available. Three of the scorers specifically praised the calibration portion of the scoring training devoted to discussion of the application of the scoring criteria. The only suggestion for improving the training was to continue refining scoring criteria to reduce the time required for scoring each task. One scorer also suggested that requiring the candidates to write legibly in dark ink might reduce the eye strain which she experienced.

The training of scorers exhibited many of the principles of good instruction. Scorers received a clear introduction to the task. Trainers monitored scorers' performance and adjusted instruction according to the results. Multiple examples were provided. The examples were randomly chosen, not chosen deliberately to illustrate different scoring decisions. For the most part, this worked well.

The training would have been strengthened by the inclusion of more examples. In similar training for assessments in the legal profession, the trainer uses fifty sample responses for calibration instead of ten to ensure variability among the sample responses. FWL staff believe that more examples would have been especially helpful for cases where the scorer was required to choose between 1, 2 or 3 points for a single appropriate item.

Assessment Content

In the following pages, the content of the Structured Simulation Tasks for Secondary Life/General Science Teachers is evaluated along these dimensions:

- Congruence with the 1990 California Science Framework;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the instrument;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching context (e.g., grade levels, subject areas);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

As was true of all of the assessment instruments pilot tested this spring and summer, there was not sufficient time during development to conduct a larger content validity study. Without such a study, our ability to comment on the assessment's appropriateness along such dimensions as job-relatedness, appropriateness for beginning teachers, and appropriateness across contexts is limited. Thus, excluding the first two dimensions of curriculum congruence and standards coverage (which are based on FWL staff's analysis of the documents involved), the discussions of the remaining dimensions are based on the perspective of the participating teachers and scorers, and FWL staff, as reflected in feedback forms, in informal conversations with the scorers and in analysis of the scores.

The discussion of the content begins with a comparison of the instrument with the preliminary edition of the 1990 *Science Framework for California Public Schools, Kindergarten Through Grade Twelve.*

Congruence with California Model Curriculum Guides and Frameworks

The California State Department of Education periodically produces subject-specific documents, curriculum guides and frameworks, which serve as public statements describing the curriculum which content and pedagogy experts believe is most appropriate for California school children. The most recent document pertaining to science is the preliminary edition of the *Science Framework for California Public Schools, Kindergarten Through Grade Twelve* (California State Department of Education, 1990 -- referred to in this report as the Science Framework). The reader should note that this framework was in development at the time of the development of the assessment, and therefore was not available to assessment developers; nonetheless, as the current statement of expectations for the California science curriculum, it will be utilized as a standard to which the assessment is compared.

The Science Framework is divided into three parts. Each part will be discussed separately, with a description of the main themes of each part followed by a discussion of whether or not the prototype tasks are consistent with the themes. FWL staff evaluations are summarized in Table 3.2.

Part I of the Science Framework discusses general characteristics of science to be emphasized in science classes, including the nature of scientific inquiry to be modeled and the thematic organization of instruction across the curriculum. With respect to the nature of scientific inquiry, two of the three lessons critiqued in the Structured Simulation Tasks for Secondary Life/General Science Teachers's **Lesson Planning** task contain student

TABLE 3.2

COVERAGE OF THE CALIFORNIA SCIENCE FRAMEWORK BY THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Content	Method of Coverage	Extent of Coverage
Part I: What is Science?		
-Nature of scientific inquiry	-Reflected in activities portrayed in all tasks.	Full
-Thematic structuring of content	-Not explicitly addressed in any task.	None
Part II: Content of Science		
-Physical Sciences	-Not represented.	None
-Earth Sciences	-One topic used for a partially developed task.	Limited
-Life Sciences	-Two topics used among all tasks.	Partial
Part III: Achieving the Desired Curriculum		
-Thinking processes emphasized	-Application reflected in the partially developed task on student misconceptions and in Parent/Student Letter.	Partial
-Level-specific guides	-Most high school level goals addressed by aspects of 1-2 tasks. No representation of middle school curriculum.	Limited
-Teaching Science to historically underrepresented students	-Some attention in Applying Effective Instructional Techniques. No representation of LEP students in any task.	Limited

practice in observation and analysis of data together with understanding why observed results need not be perfectly consistent with predicted results to support a theory. The rest of the tasks, with the possible exception of **Teacher as Curriculum Decision-Maker**, address different aspects of science instruction, and there is a consistent use of hands-on activities to illustrate concepts being taught throughout the assessment.

The emphasis on thematic structuring of content in the Science Framework addresses the use of themes both across courses and within a course. With respect to the former, this specific kind of articulation of course content both with previous science courses at earlier grade levels and with other science courses at the same grade level would be difficult to illustrate within a unit, much less through a single lesson. It is also questionable if all beginning teachers could be expected to articulate course content at such a grand scale. However, the thematic structuring within a course could be reflected in the present tasks with slight modifications of the contextual information and stimulus materials. Information on the theme(s) emphasized could be included in the contextual information for all tasks and the extent to which instruction reflects the given theme(s) could be incorporated into the scoring criteria for **Lesson Planning** and **Teacher as Curriculum Decision/Maker**. In addition, the current tasks all focus on high school classes. Since high school classes tend to be more specialized, forms of tasks which address middle school classes might exhibit the thematic emphasis more clearly.

Part II discusses specific content to be presented at different grade levels and how it might vary in presentation according to the themes emphasized. The topics chosen for representation in the assessment could be more diverse. The tasks pilot tested focus on only two of the fifteen topics described under curriculum content, both within the life science curriculum: "Living Things" and "Cells, Genetics and Evolution." (An overrepresentation of topics from the life sciences is to be expected, as the assessment covers both life science and general science; physical science teachers would take a separate assessment.) In addition, one partially developed task involved classification of minerals, part of "Geology and Natural Resources".

Some of the content of lessons portrayed, e.g., the lesson in **Applying Effective Instructional Techniques**, seems to be more characteristic of the instruction called for at earlier grade levels. One complication is that the content called for in the Science Framework represents a model to work toward and does not reflect the content presently taught at specific grade levels, especially at the elementary level. Thus, until the content taught is more in alignment with the State Science Framework, a policy decision may be needed as to the extent to which a secondary science assessment should reflect either the grade-level content in the Science Framework or the content typically taught at that level. Whatever the ultimate decision, the present task shells could easily be modified to reflect a greater variety in representation of content as well as the content deemed appropriate for secondary science.

Part III discusses implementation of the desired curriculum. While the chapters discussing programmatic implementation at the school district and site level and criteria for the adoption of instructional materials are distant from the responsibilities of typical beginning teachers, the chapter on "Science Processes and the Teaching of Science" describes desirable characteristics of science instruction which are applicable to the classroom. These characteristics include thinking processes to be emphasized, guides for science programs across grade levels, and an emphasis on teaching science to the historically underrepresented (females, most minority groups, and the disabled) and Limited-English-Proficient students.

The Science Framework calls for an emphasis on the thinking processes of observing, communicating, comparing, ordering, categorizing, relating, inferring, and applying. All thinking processes are not to be taught in all grades, however, as theories of child development suggest that young children develop these skills sequentially, and roughly in the order listed. While all antecedent skills are reinforced and refined at all levels, inferring is to be introduced and stressed in grades six through nine, and applying in grades nine through twelve.

As all tasks concentrated on high school classes, the skill of applying is most pertinent. The Science Framework discusses this skill in the context of learning to use scientific knowledge to think about current problems. The only task in which this is specifically done is the task examining students' scientific misconceptions, where the students discuss a current problem in light of the science which they have just learned. In addition, scoring criteria for the task **Parent/Student Letter** imply that students will learn to apply their scientific knowledge to current problems.

Activities portrayed in the tasks pilot tested included several instances where the students were required to use the antecedent skills, particularly observing when conducting experiments in **Lesson Planning** or **Teacher as Curriculum Decision-Maker**, comparing as when two types of cells are contrasted in **Teacher as Curriculum Decision-Maker**, and inferring during activities analyzed as part of **Lesson Planning**. Task shells could easily be revised to include a focus on reinforcing and/or developing various types of thinking skills, and a variety of thinking processes could be represented across the set of tasks which compose an assessment.

The Science Framework provides level-specific guides for science programs. For secondary science instruction, both middle school level and high school level guides are provided. However, no middle school level courses were portrayed in the current forms of tasks developed. The goals emphasized for high school science programs in the Science Framework are shown in the following bulleted paragraphs in italics, followed by a description of the extent to which the assessment reflects each goal.

- *Build on a solid foundation of science instruction in kindergarten through grade eight.* At present, elementary schools are in various stages of aligning the content of their science instruction with previous Science Frameworks. However, most elementary schools do not yet teach all the content described in the 1990 Science Framework, so content previously taught in elementary grades would need to be specified in any task which addressed this aspect of science instruction. At present, no tasks specify the science content in the elementary curriculum which teachers should assume the students experienced.

- *Lead in a coherent fashion to greater opportunities for all students.* This goal stems from a desire to make science comprehensible to a wider range of student, especially students whose limited mathematical experience may have prevented them from meeting prerequisites for science courses. It also calls for more integration of the science curriculum and less discipline-oriented courses which emphasize the common foundation of basic principles of physics, chemistry, and biology.

  The collection of tasks pilot tested portrayed a number of different classrooms; however, most of the tasks focus on students who do not plan to go to college, a group which in the past would have been less likely to take science courses than college-bound students for whom science courses have been required. One of the tasks pilot tested presumes that the students in the classes for which the teacher is performing the task is evenly split between college-bound and non-college-bound students; three tasks focus on classrooms of non-college-bound students; and the remaining task focuses on a classroom of mostly college-bound students. Both **Lesson Planning** and **Teacher as Curriculum Decision-Maker** contain scoring criteria which focus on the teacher's ability to recognize specific aspects of the lessons or activities which are especially appropriate or inappropriate for most non-college-bound students. Teachers reported difficulty with this aspect of the assessment; this will be described in detail later in this section of the chapter.

With respect to integration of the science curriculum, all the tasks focus on classes at the high school level, which would be expected to be more specialized than those at other grade levels. The Science Framework cites examples of integration such as a biology class examining the physics of motion and the concept of work and machines when discussing bones and muscles. FWL staff finds no such instances when other scientific disciplines are integrated in the tasks that were pilot tested.

- *Help students understand the nature of science -- in particular, its experimental, nondogmatic nature and the methods by which progress is made.* This is the nature of scientific inquiry previously discussed with respect to Part I of the Science Framework. The activities portrayed in the tasks and the scoring criteria are consistent with this emphasis.

- *Develop in students a strong sense of the interrelationship between science and technology and an understanding of the responsibility of scientists and scientifically literate individuals to both present and future societies.* No task pilot tested specifically reflects this goal. One of the scoring criteria for a partially developed task addressing using computers as tools in science education is whether or not the teacher recognizes that a weakness of the lesson portrayed is the missed opportunity to link the lesson to career options and relevant real-world uses of data bases.

- *Foster each student's ability to act as an independent investigator and thinker rather than a "recipe follower."* **Lesson Planning** explicitly includes this as one of its scoring criteria for the lesson developed by the candidate, i.e., that the lesson is not merely rote learning. Many other criteria for judging the analysis of lessons provided in **Lesson Planning** are focused on developing this ability in students. A few examples of relevant scoring criteria include recognizing the improvement of students' problem solving skills, practice in data collection and analysis, the manipulation of materials, and the demonstration of several scientific principles (e.g., that more samples lead to more valid results) as strengths and not involving enough students directly in the activity and asking too many similar questions in the activity as weaknesses.

- *Reinforce basic tools of language and mathematical communication.* This goal calls for more integration across subjects so that students receive reinforcement for writing and mathematical skills in classes other than English and mathematics, and practice writing and mathematical problem-solving with topics which would normally be found in other classes, such as science. Although instances where the student communicated in writing or needed to use certain mathematical skills were portrayed in the tasks pilot tested, the specific emphasis portrayed in the Science Framework was not specifically reflected in any task pilot tested.

- *Provide an expanded view of science-related careers.* This was not addressed in any of the tasks pilot tested. However, one of the scoring criteria for a partially developed task addressing using computers as tools in science education is whether or not the teacher recognizes that a weakness of the lesson portrayed is the missed opportunity to link the lesson to career options.

There was some attention to historically underrepresented students in the task of **Applying Effective Instructional Techniques**, mainly in equitable and non-racist instruction and encouragement of students. The Science Framework also suggests providing diverse role models, providing extracurricular enrichment opportunities, building parent involvement and peer recognition programs, and building on prior student knowledge to either draw on or augment student background knowledge.

Instruction of Limited-English-Proficient students was not addressed in any of the tasks.

While the present collection of tasks only partially covers the specific emphases in the latest Science Framework, none of it is in contradiction to the framework. Task shells could easily be modified to cover a larger portion of the Science Framework.

Extent of Coverage of California Standards for Beginning Teachers

The California Beginning Teacher Standards are criteria for teacher competence and performance which the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. Listed below are brief italicized descriptions of Standards 22 through 32 which pertain to expectations of student competencies to be attained prior to graduation from teacher preparation programs. (The remaining standards address programmatic requirements.) To evaluate this assessment instrument and make inferences about the assessment approach which it represents in terms of the appropriateness for use with California secondary life and general science teachers, the stimulus materials and scoring criteria for each task were compared with the 11 California Beginning Teacher Standards. Each standard will be discussed separately.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* Although this was not measured directly, in **Applying Effective Instructional Techniques,** teachers need to identify appropriate and inappropriate teacher interactions with students. These include both appropriate and inappropriate responses to students, appropriate and inappropriate use of discipline, and instances of inequitable treatment of students and racially insensitive remarks. Because this skill includes a teacher's interpersonal and group management skills, it is difficult to simulate through a transcript. Its complete measurement probably relies upon direct observation.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* These skills are at the heart of the **Teacher as Curriculum Decision-Maker** exercise where a teacher needs to be able to choose and sequence activities into a unit of instruction which meets the learning objectives, includes differing approaches to learning, and uses classroom time efficiently. The teacher also chooses an appropriate assessment activity as part of this exercise. The **Lesson Planning** task also requires these skills. Even though teachers only evaluate single lessons, the evaluation includes whether each lesson contributes toward meeting the unit objectives, correctly presents content which is appropriately sequenced in relation to previous and subsequent lesson topics, and includes appropriate activities for the grade level and achievement level of the students. The aspects of curricular and instructional planning mentioned in the standard are important contributors to a teacher's score on these two tasks, **Teacher as Curriculum Decision-Maker** and **Lesson Planning**.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* Both the **Teacher as Curriculum Decision-Maker** and **Lesson Planning** tasks evaluate whether the activities and materials chosen by the teacher are appropriate to the specific group of students described in the contextual information provided. Moreover, in **Teacher as Curriculum Decision-Maker**, teachers are evaluated on whether they include a variety of activities among those provided. This variety could be constructed to represent diverse learning styles. In the introductory material for each task, students are described in terms of their grade level and sometimes plans for education beyond high school, but not in terms of interests or learning styles. It would be possible to slightly revise the tasks mentioned to more completely address this standard by including student interests and learning styles in the contextual information provided, revising the activities provided to the teacher to include both appropriate and inappropriate activities given the interests and learning styles described, and adding scoring criteria which evaluate the match between chosen activities and the students described.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* In the task **Applying Effective Instructional Techniques**, teachers are asked in several instances to identify where a teacher uses appropriate or inappropriate techniques to motivate, involve, or discipline students. However, responding to a transcript of classroom interactions only captures limited features of the complex task of motivating and equitably sustaining student interest, involvement and appropriate conduct.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* The **Applying Effective Instructional Techniques** task contains several instances in which teachers need to identify either appropriate or inappropriate representations of concepts or structuring of the lesson. In **Lesson Planning**, teachers are asked to describe a lesson that they would teach. Part of the criteria by which this lesson is judged is the extent to which the concepts relate to objectives, are appropriately sequenced from easy to more complex, are reflected in classroom activities and homework, and are scientifically correct. A task which was not fully developed addressed student misconceptions of scientific concepts, and could easily contain a component asking the teachers to describe how they would explain a key concept.

Any aspect of this standard which addresses such performance aspects of presentation as whether a teacher speaks loudly and clearly enough to be understood, however, would be difficult to capture with pencil and paper tests such as the Secondary Life/General Science Assessment.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* Of the tasks piloted, structuring of lessons to achieve significant instructional objectives was addressed by **Teacher as Curriculum Decision-Maker** and **Lesson Planning**. The **Teacher as Curriculum Decision-Maker** task includes selection of an appropriate summative evaluation instrument for the class described.

None of the tasks pilot tested address student diagnosis or evaluation. However, two other partially developed tasks address this standard. Student diagnosis is the focus of a task which focusses on common student misconceptions about phenomena which conflict with accepted scientific theories. The task requires teachers to evaluate the strengths and weaknesses of instruction embodied in a script of teacher/student interaction which exhibits student misconceptions. Teachers diagnose or evaluate the student responses, analyze two possible scenarios of remedial instruction, and outline their own lesson(s) to correct the student errors exhibited.

Another task focuses on the evaluation of student performance. The task requires a candidate to (1) determine if a set of instructions is clear or what modification need to be made to make the assignment clear; (2) determine if responses by another grader are accurate and appropriate; and (3) list common factual errors in a group of student papers and design a lesson to correct those misconceptions.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* Even though cognitive outcomes are not addressed, the evaluation of lessons designed by the candidate in **Lesson Planning** includes whether each lesson contributes toward meeting the unit objectives, correctly presents content which is appropriately sequenced in relation to previous and subsequent lesson topics, and includes appropriate activities for the grade level and achievement level of the students. The incompletely developed task which requires teachers to diagnose students' scientific misconceptions also measures their ability to recognize effective remedial strategies which would foster the ability of all students to evaluate information, think analytically, and reach sound conclusions.

Standards 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* **Applying Effective Instructional Techniques** contained several instances where a teacher needed to identify appropriate or inappropriate actions which would affect student attitudes toward science and/or other students. **Lesson Planning** addressed some aspects of fostering a student's capacity to become an independent learner, i.e., sequencing concepts within the lesson from easy to more complex and assigning homework of appropriate difficulty which is not rote.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socioeconomic differences.* One segment of **Applying Effective Instructional Techniques** contains several inappropriate teacher remarks to students which are ethnically or culturally insensitive. No other issues of diversity listed in the standard, i.e., gender, linguistic, socioeconomic, were explored.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* Although the shells for the task include classroom context as one of the variables to be manipulated, the set of tasks piloted focus on high school (mostly tenth grade) classes. There is some diversity in terms of ability levels of students, and teacher responses need to take ability level into account in most of the exercises. There is no reason why the ability levels and school contexts could not be varied more widely if the assessment approach were used for credentialing.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interaction with other members of the profession.* Although the **Parent/Student Letter** addresses communication with adults (i.e., parents), it focuses on one small aspect of cooperation with parents, informing them of the learning objectives and content of a particular science course, and general classroom policies. The scoring criteria for the letter focus on the topics included in the letter, and not on whether the letter is comprehensible to the parent population. Moreover, the language used in the example in the instructions would be difficult to comprehend for most adults who are unfamiliar with the technical terms used, much less for Limited-English-Proficient parents:

> *A topic might be "lab safety" and an example would be: "student will wear approved protective goggles when doing laboratory experiments where a splash hazard exists."*

The only other task depicting other adults in the school community is **Classroom and Facility Safety**, where teachers draft a memo concerning a specific safety issue to be distributed to all science teachers in the district.

Table 3.3 lists the standards and FWL staff's evaluations of the extent to which the assessment methodology covers each standard, based on reviews of the fully and partially developed task shells together with a consideration of alternative tasks that might easily be developed. A "full" rating indicates that multiple dimensions of the standard impact a teacher rating, even if these dimensions are not scored separately. A "partial" rating indicates that some dimensions affect the rating, but some important dimensions are unexamined. A "limited" rating indicates that some dimensions affect the rating, but most important dimensions are unexamined.

The previous two sections have addressed the congruence of the assessment with state standards. To address other aspects of the content, teachers and scorers completed surveys soliciting their perceptions of the appropriateness of the assessment along a number of dimensions: job-relatedness, appropriateness for beginning teachers, appropriateness across contexts, fairness across groups of teachers, and, finally, a general evaluation of the appropriateness of this method of assessment. Their perceptions, together with data on teacher performance, are summarized in the remainder of this section.

Job-Relatedness

Both teachers and scorers were asked whether the tasks chosen were relevant to the job of teaching life/general science at the secondary level.

Teacher perceptions. Teachers agreed that the tasks were relevant to their job of secondary science teacher. Approximately 92% (12 of 13; the relevant page was missing from the survey form for 19 teachers) of the teachers responding to the question who completed form A and 76% (25 of 33) of those completing form B responded affirmatively.

TABLE 3.3

EXTENT OF COVERAGE BY THE SECONDARY LIFE/GENERAL LIFE SCIENCE TEACHER ASSESSMENT OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Tasks Addressing Standards	Extent of Coverage
22: Student Rapport and Classroom Environment	-Applying Effective Instructional Techniques	Partial
23: Curricular and Instructional Planning Skills	-Tchr as Curric Decision-Maker -Lesson Planning	Full
24: Diverse and Appropriate Teaching	-Tchr as Curric Decision-Maker -Lesson Planning	Partial
25: Student Motivation, Involvement and Conduct	-Applying Effective Instructional Techniques	Limited
26: Presentation Skills	-Applying Effective Instructional Techniques -Lesson Planning	Partial
27: Student Diagnosis, Achievement and Evaluation	-Tchr as Curric Decision-Maker -Lesson Planning -Student Misconceptions -Evaluation of Student Performance	Full
28: Cognitive Outcomes of Teaching	-Lesson Planning -Student Misconceptions	Full
29: Affective Outcomes of Teaching	-Applying Effective Instructional Techniques -Lesson Planning	Limited
30: Capacity to Teach Crossculturally	-Applying Effective Instructional Techniques	Limited
31: Readiness for Diverse Responsibilities	-Partial	Partial
32: Professional Obligations	-None	Limited

Few teachers elaborated on their response; those teachers who singled out a single task for praise mentioned **Classroom and Facility Safety**. The positive responses included:

> *I really enjoyed the tasks. They were very appropriate.*
>
> *Subject content: Not more specifically, because I teach in a middle school, and the time allotment is much less for each topic. Everything else: yes.*

Of those teachers who did not believe that the tasks were relevant, their criticism mostly addressed the content embedded in the tasks, and not the tasks themselves, as illustrated by the following comments:

> *I felt the lesson planning section, the examples, were not appropriate for the particular grade level I teach. It should be more grade oriented.*
>
> *I will touch only briefly in genetics and if it is covered in tenth grade classes that is fine, but I teach 7 grade Life Science and it does not get into that much depth.*
>
> *At my school site, we have a safety coordinator who is in charge of storing chemicals, etc. Therefore I have no contact with most of the chemicals displayed.*

Some teachers, however, did question the ability of the specific tasks to measure their teaching skills, as illustrated in the following quote:

> *Yes! The tasks are relevant...I do employ instructional techniques...I do lesson planning... I maintain classroom safety. However...This test does little to accurately assess my skills in these areas (Except for the Classroom Safety -- Section 3B -- it is fine).*

Scorer perceptions. All four scorers agreed that the tasks were relevant for secondary science teachers. One scorer expressed concern about the ability of non-inner city teachers to catch the subtlety of the dialogue in **Applying Effective Instructional Techniques**. Another scorer who participated in designing the assessment thought that the assessment could serve as a useful tool for shaping the curriculum in teacher training institutions: "If weaknesses in candidate responses are throughout the assessment, then the instrument should drive the programs in universities and districts."

Generally, both the scorers and the teachers believed that the tasks chosen were relevant to the job of secondary science teachers. For the most part, those who disagreed believed that the content contained in the task, and not the task itself, was inappropriate.

Appropriateness for Beginning Teachers

Because the focus of this assessment is on beginning teachers, who are still in the initial stage of professional development, one issue pursued was whether or not the tasks were too easy or too difficult.

Perceptions. When asked whether they had "sufficient opportunity to acquire the knowledge and abilities needed to respond in a reasonable manner to the assessment questions," 92% (12 of 13) of the teachers responding to the question with respect to Form A and 70% (23 of 33) of the teachers completing Form B responded affirmatively. Some teachers who believed themselves well prepared went on to emphasize that their experience teaching played a critical role in their preparation:

> *As a second year teacher, I feel that I have developed some of the skills needed to take this test. But I have done so only because I have survived two very stressful years in teaching.*
>
> *Yes, but only because I have experienced a full load of teaching and worked on some of the frustrations, problems, etc. for at least one year. I do not think I would have been as well prepared after my student teaching because (1) no specific attention was given to safety; (2) my students for student teaching were very good, motivated, etc. I may not have been as aware of the need to build esteems and motivate if I had not taught in two other classrooms after student teaching.*

Teachers who did not feel well-prepared gave a number of explanations. Some teachers mentioned the topic:

> *I felt very uncomfortable trying to write a lesson on genetics when I have never taught genetics. Could this be a fair evaluation of my competency?*
>
> *I was a geology major! My biology background is very weak.*

Some mentioned a perceived need for more or more varied classroom experience:

> *I believe a lot of the assessment questions require being in the classroom for a few years and gaining the experience to know what is appropriate and what is not.*
>
> *Knowing about weaknesses and strengths in lesson planning comes with a lot of experience and trial and error.*
>
> *Abilities, yes. Not necessarily all of the knowledge -- a lot of that you pick up from the specific course(s) you teach.*

Finally, some criticized the tasks themselves:

> *Many unanticipated and unexpected things happen in a classroom and they are difficult to anticipate and, therefore, difficult to assess methods to prevent or eliminate them.*
>
> *I don't think there is a reasonable response to the assessment questions. They are very poor. The only assessment questions that were reasonable were the safety ones.*
>
> *But the questions, being out of context, are absurd.*

When the teachers were asked if they found any tasks or parts of tasks too difficult, 38% (12 of 32) of those completing Form A and 42% (14 of 33) of those completing Form B replied, "Yes." Nine teachers identified **Teacher as Curriculum Decision-Maker** for reasons such as the following:

> *There appear to be too many objectives to cover in a two week period in the Teacher as Curriculum Decision-Maker exercise.*

> *It was more of a jigsaw puzzle -- moving pieces of paper around, trying to add up to 50 minutes. I am much more flexible in my planning.*
>
> *Using* *only* *those activities listed to develop a two-week curriculum. For example, I wanted to give a quiz after the first week, but I did not like the specific content of the only quiz listed -- in terms of correlating with my first week's content.*
>
> *All were difficult from the point of view that most of the situations are out of context. I.e., [SES] of class, type of school, climate of classroom and hundreds of other variables. Although some of the situations seemed cut and dry, others were subjective relating to the above variables.*

For Form B, 13 of the 33 teachers identified **Lesson Planning**, 3 **Applying Effective Instructional Techniques**, and 2 **Classroom and Facility Safety**. Teachers did not typically give lengthy explanations of their choice. The teachers selecting **Lesson Planning** mentioned: (1) the topic, as in "In 7th grade we do not go into a three week course in genetics;" (2) difficulty in understanding the kind of response expected, as in "Lesson planning wasn't difficult, but I didn't know how much to write, how thorough to be;" and (3) technical problems, as in "Having to come up with a lesson plan without having knowledge of what was specifically taught previously." The few teachers who found **Applying Effective Instructional Techniques** difficult either felt that it was too "taxing" to identify and explain both appropriate and inappropriate actions or felt that it was difficult to evaluate out of context, preferring to see either a real teacher or a video. The teachers choosing **Classroom and Facility Safety** found critiquing the drawing depicting the storage of chemicals difficult.

The scorers evaluated the difficulty of the tasks from a different perspective, having seen both the expected answers and the teacher responses. According to the scorers, teachers had problems with three tasks in particular: **Parent/Student Letter, Classroom and Facility Safety, and Lesson Planning**. Many teachers seemed to have no experience in writing parent/student letters, exhibiting a lack of ideas of what might be included. With regard to safety, one scorer remarked, "Many teachers seemed unprepared to deal with the classroom situation and the storage area. This is so important, it should be on all formats and carry weighted points. Once the teachers and prep program advisors know there is accountability, the performance level will improve." Both scorers of **Lesson Planning** commented on the lack of expertise exhibited by the teachers, described by one scorer as "abbreviated, shallow candidate answers resulting in very low scores." This scorer went on to suggest that perhaps candidates felt rushed, yet only one teacher completing this task reported a need for more time while 1/3 of the teachers (11 out of 33) identified it as being too difficult.

While the teachers for the most part believed that they had an opportunity to develop the skills required of the tasks, the scorers disagreed. On numerous occasions during the scoring, various scorers commented on what they perceived as the inadequacy of current teacher training programs based both on their experience with new teachers and the answers being scored. The scorers expressed hope that an assessment such as this one would provide guidance for curriculum development and feedback on the performance of the graduates of teacher preparation programs.

Performance on assessment tasks. Table 3.4 shows a statistical portrait of teacher performance on the assessment tasks. (Both the scores and the number of possible points were doubled, as scores were formed by adding the scores of the two scorers.) Teacher scores suggested that the content was difficult for teachers. Teachers as a group did best on **Teacher as Curriculum Decision-Maker** and **Parent/Student Letter**. The lowest scores were recorded for the portions of **Lesson Planning** that required teachers to analyze another teacher's lesson plan.

To do well on a task, a teacher had to pay attention to many simultaneous factors and attend to most of the cues provided, e.g., lesson objectives, classroom composition. While the multiplicity of cues reflects the complexity of classrooms, the burden is on the teacher to process a great deal of information and to place the same significance on certain cues as the assessment developers, who were recognized experts on science teaching.

Teachers reported difficulties when faced with unfamiliar situations. This appeared across tasks, as when teachers reported unfamiliarity with the topic in **Lesson Planning** or when the teaching approach they commonly used was not congruent with the activities provided in **Teacher as Curriculum Decision-Maker** task or when they had never seen, let alone written, a parent/student letter. Teachers also seemed to have trouble with designing instruction for a group of students who were different than their own. This was reflected not only in relatively low scores and what one scorer called "shallow" responses, but also in a common criticism of the activities in the tasks as inappropriate for their own students, despite clear directions that they were to plan for a different group of students. The scorers/assessment developers were confident that new teachers could be trained to design

TABLE 3.4

TEACHER PERFORMANCE, BY SUBPART, ON THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

	Teacher Scores*			
Subpart	Mean	Standard Deviation	N	Points Possible
Form A				
Task 1: Applying Effective Instructional Techniques				
Segment 1	4.8	3.0	32	18
Segment 2	7.3	3.3	32	20
Segment 3	7.9	2.7	32	14
Segment 4	7.6	3.1	32	14
Task 2: Teacher as Curriculum Decision-Maker				
Part I	113.9	20.9	32	200
Part II	17.8	8.5	32	33
Task 3: Parent/Student Letter				
Part I	9.8	4.1	32	16
Part II	17.7	10.9	32	30

*Since each teacher response was double-scored, the scores were derived by summing the two ratings.

TABLE 3.4 (Continued)

TEACHER PERFORMANCE, BY SUBPART, ON THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Subpart	Teacher Scores* Mean	Teacher Scores* Standard Deviation	Teacher Scores* N	Points Possible
Form B				
Task 1: Applying Effective Instructional Techniques				
Segment 1	5.3	3.3	33	18
Segment 2	7.7	3.3	33	20
Segment 3	4.0	3.7	33	14
Segment 4	15.7	5.6	33	42
Task 2: Lesson Planning				
Part I: Lesson A	4.8	2.4	33	30
Part I: Lesson B	4.5	3.3	33	26
Part I: Lesson C	2.2	1.6	33	14
Part II: Organization	4.0	1.9	33	8
Part II: Content	3.2	1.5	33	8
Task 3: Classroom/Facility Safety				
Part I	13.1	5.3	33	28
Part II	10.2	3.9	33	42
Part III	15.2	6.2	32	52

*Since each teacher response was double-scored, the scores were derived by summing the two ratings.

instruction for different groups of students, but research has identified this as an area which distinguishes new teachers from expert teachers (Leinhardt, 1983; Wilson, 1988), suggesting that this may be a skill that develops later in a teaching career. New teachers generally have experience with a very limited range of students, those who they taught during student teaching and those taught during one or two years as a regular teacher. Can they effectively build on the more in-depth experience as a full-time teacher to think about what may be appropriate for different types of students or are they so caught up in classroom management, time management, and lesson planning that the issue of tailoring instruction develops later? This question cannot be answered by any data we have, but is key to ascertaining the appropriateness of the assessment for beginning teachers.

Appropriateness across Contexts

The assessment is designed so that the teaching context can be varied. The tasks piloted were very homogeneous with respect to grade level. All tasks focussed on high school classes consisting mainly of tenth graders. However, 54% of the teachers taking the test taught students who had not yet reached the ninth grade; two of these teachers taught sixth grade in a middle school with their single subject science credential.

The tasks were more heterogeneous with respect to students. One task, **Teacher as Curriculum/Decision-Maker**, featured students who were not planning to attend college. These students were described in the following way: "Though students read at grade level, they are not used to having extensive assignments and many do not even complete minimal homework." **Classroom and Facility Safety** and **Parent/Student Letter** featured a class of students who were non-college bound, ranging from ninth to twelfth grade, though containing mostly tenth graders. **Evaluating Effective Instructional Techniques** focussed on a class fulfilling a college entrance requirement, but did not specify the kinds of students in the focal classroom. **Lesson Planning** portrayed a class where half the students planned to attend college, and half did not. **Classroom and Facility Safety** did not specify the type of students, although the class itself is described as non-college preparatory.

Teachers were asked their perceptions of the appropriateness across contexts on two dimensions: with respect to teachers at different grade levels, and with respect to teachers of diverse types of students.

Grade level. Teacher perceptions of the appropriateness of the assessment across contexts differed according to the form completed. Teachers completing Form A overwhelmingly (75% or 24 of 32 teachers) agreed that the assessment was appropriate across grade levels. Only 48% (16 of 33) of the teachers completing Form B, however, agreed. Dissenting teachers completing either form, however, tended to agree that they saw the assessment as problematic for junior high and middle school teachers, as exemplified by the following comments:

> *I think junior high teachers would have a hard time with appropriate high school level activities which test asks for.*
>
> *Lesson Planning was too in depth for 7th grade advance ESL science class. We have to break down the lesson step-by-step due to the language problem that sometimes occurs.*
>
> *Teachers teaching junior high school have more to deal with as far as student achievement levels.*
>
> *The materials given for the unit planning are too advanced for junior high students. Although, teachers are supposed to be able to cover all grade levels even though they may teach at a different level from that which they were assessed.*

Some junior high teachers in response to earlier questions also remarked that they did not cover topics in the depth required by the **Lesson Planning** and the **Teacher as Curriculum Decision-Maker** tasks.

Scorers did not comment, either positively or negatively, on the grade-level aspects of the instrument's appropriateness.

Diverse students. Teachers were asked whether they felt that the assessment was "appropriate for science teachers of diverse student groups (e.g., different student ability levels, different ethnic groups, handicapped or Limited-English-Proficient students, different school/community settings." Sixty-nine percent (22 of 32) of the teachers completing form A and 56% (19 of 33) of the teachers completing form B believed that it was appropriate. These teachers supported their response with comments such as the following:

> *It is still teaching.*
>
> *In all but extreme cases.*
>
> *However, more diverse examples could be used.*

Teachers who disagreed generally cited types of students who were not represented in the stimulus materials or the generalizability of the settings used.

> *Parts I and II both were biology. Many students do not take Biology.*
>
> *The lesson plans and self directed homework assignments are much too difficult for the students of my district. I have LEP and PL students who wouldn't be able to handle the written work -- especially the math.*
>
> *I'll use my district as an example. Most of my parents don't speak English. My department has no equipment or budget for the labs described and no films (have to order one year in advance -- impossible for a first-year teacher).*

It is not clear whether the teachers who criticized the assessment as including teaching techniques and conditions which were inappropriate for the students they taught realized that some context was provided in the instruction for the tasks and believed that more diverse contexts should be represented, whether they believed that their teaching context should have been represented for it to be a valid assessment of their teaching, or whether they missed the contextual remarks in the introductory materials for the tasks.

Teachers are licensed to teach all students, so it seems reasonable to present varied groups of students in the stimulus materials to test whether or not a teacher knows how to vary instruction. However, this skill is known to be more characteristic of experienced teachers than beginning teachers (Leinhardt, 1983), so beginning teachers may have difficulty in completing tasks for students with whom they have limited or no experience. One teacher who did feel that this assessment was appropriate for teachers of differing student groups summed up the dilemma of addressing teaching diverse students in an assessment:

> *Teachers that have different ability levels and handicapped and limited English students should be tested for this. But wait...all teachers will probably be exposed to these types of students and should be tested on their ability to handle the problems that could arise. But...I have not been trained to handle these students and have trouble finding answers considering the lack of resource specialists, resources, materials, textbooks, and the ratio of students to teachers. In my school it is 34:1 on average! Tough question!*

All of the four scorers felt that the assessment was "suitable for new teachers in different school and community groups." Opinions ranged from "Regardless of the school and community setting, all teachers need to be aware of the components in each and all of the tasks" to "There is some potential for [being unsuitable]. However, I can think of no way to prevent this. A broad-based series of questions should not penalize any one type of teacher too much."

Another perspective on the appropriateness of the tasks for teachers of diverse students was obtained through review of portions of two tasks by Dr. Sharon Nelson-Barber of Stanford University, a consultant who works with school districts and teachers of classrooms composed primarily of students outside of the dominant culture. The materials sent included both stimulus materials and scoring criteria for one segment of **Applying Effective Instructional Techniques** and materials from both parts of **Lesson Planning**, including one of the lessons to be critiqued and the lesson to be designed.

Dr. Nelson-Barber praised the provision for rater recognition of appropriate responses which are not included on the scoring guide. However, she emphasized the need for test developers to consider a variety of perspectives as the scoring criteria are devised. One example she cited was the literature on effective black teachers' emphasis on strong adult leadership (Hollins, 1982; Delpit, 1988; Foster, 1989; Ladson-Billings, 1889) as contrasted with more mainstream characterizations of good teaching as guiding and facilitating, i.e., deemphasizing the authority role. As an example of potentially conflicting notions of effective teaching, Dr. Nelson-Barber cited as an example an instance of teacher sarcasm in the transcript analyzed that was evaluated as an inappropriate teaching behavior. However, certain culturally sanctioned teasing behaviors or "put downs" built upon shared backgrounds and cultural understandings between teachers and students have been used very effectively with black inner-city college students (Foster, 1989). It is likely that any similar teacher responses that lie outside the scorer's range of cultural experience and/or knowledge, but which may represent culturally appropriate and highly effective practice within the context of the teacher's particular teaching context will be either ignored or negatively evaluated. One way to reduce the likelihood of this occurring is to require review of an assessment by a number of successful teachers working in culturally diverse settings.

Fairness across Groups of Teachers

Teachers were asked whether or not they felt this assessment was "fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers." Teachers overwhelmingly believed that the assessment was fair to different groups of teachers. For those teachers completing Form A, 91% (29 of 32) of the teachers agreed that it was fair; for teachers completing Form B, 85% (28 of 33) agreed.

One teacher supported her affirmative answer by noting the diversity of students in the prompt materials. Another teacher who felt that the test was fair believed that "these factors should not be a concern."

Teachers who disagreed gave differing reasons. Three teachers did not cite specific groups for whom they believed the assessment to be unfair, but instead expressed their disapproval of the entire assessment and their belief that no one should have to take it.

Two other teachers expressed concern for teachers of varying English proficiency or cultural backgrounds:

> *The language/culture differences of teachers are not addressed by the wording of the test. Someone who is not fluent in English may have difficulty with some terms. Also, some cultures may take a more regimented view of classroom management.*

Two other types of teachers elicited the concern of teachers:

> *I am a bilingual science teacher which should be assessed along with other mainstream classes I also teach.*

BEST COPY AVAILABLE

> *Some teachers who were taught by the university system may have gotten more experience when it came to writing lesson plans than teachers who were taught by other alternative means.*

Finally, one teacher believed that the test was fair only if "teachers are allowed to choose the area to be tested in, for example, 'Cell Theory'." This comment mirrored the frustration of other teachers who described themselves as junior high school general science teachers faced with designing a laboratory in genetics.

Three of the four scorers believed that the assessment was fair to all groups of teachers. As one scorer commented, "This assessment is directed to the skills needed to teach California public school science students. Teachers with various characteristics and teaching styles must be at least minimally proficient in the needed basic science teaching skills." The fourth scorer believed that "If the teacher does not have good command of English, this will be a problem." Whether or not "good command of English" was a necessary prerequisite for good science teaching was not addressed by this scorer.

The expert on teaching diverse students, Dr. Nelson-Barber, stressed the need for specifying the information that the candidate is expected to provide for each task. For instance, many members of the black community, particularly working class blacks, use a communicative style that devalues the expression of "obvious" information (Heath, 1983; Taylor and Lee, 1987). In responding to an assessment, a teacher may not display the full range of their knowledge because aspects considered to be "obvious" are not mentioned. Epistomological or communicational patterns from other cultures may present additional problems.

Appropriateness as a Method of Assessment

Teachers were asked directly whether or not they thought "this type of assessment is an appropriate way of assessing your competency in teaching secondary life and/or general science." About 77% (9 of 13; this question was inadvertently omitted from the surveys of 19 teachers) of those responding to the question for form A and 52% (17 of 33) of those completing form B believed the assessment to be appropriate.

Teachers responding positively had comments such as the following:

> *I think this is a good start -- I think it assesses knowledge of teaching skills more than subject content.*
>
> *If the bugs are worked out and a realistic way to grade this mass of paperwork is discovered!*
>
> *My answer is yes and no because this could be one way of testing for competency. I feel classroom observations are important also. Actual teaching and a written assessment are two totally different things.*
>
> *More appropriate than the kind of assessment I received in my credentialing program. There was not check for competency other than knowledge of subject and pre-arranged classroom observations.*

Teachers who did not feel that the assessment was appropriate for measuring their competence offered specific criticism, perceiving a lack of measurement of collegial and interpersonal interaction variables, a need for more measurement of teaching culturally diverse students, a need for greater emphasis on cooperative learning, and general skepticism about the ability of pencil-and-paper tests to indicate teaching competency. The following is a sample:

> *It's not at all close to what it's like to be a teacher. For example, when I send letters home to parents I always consult other teachers as I do when I plan a unit, etc. So much of being a good teacher depends on communicating with other teachers, with your personality around students, etc.*
>
> *Partially so, more emphasis should be placed on the assessment of culturally aware and sensitive teachers and their use of cooperative learning! No more book/lecture teaching!!*
>
> *There are so many variables in teaching. It is preposterous to think that a pen-paper test (in which I evaluate a script of an awful lesson or I develop a particular lesson for a particular subject for a particular grade for UNKNOWN students, sites,*

> *materials, staff, etc.) will in any way be an indicator of teaching competency.*
>
> *In motivating students -- it is not only science, science, science, to build relationships, the teacher needs to be a little more personal. Part III-Safety -- not all science teachers [go] into the stockroom with chemicals. There are separate stockrooms for life and physical [science].*

Comparison with other assessments. Teachers were also asked the following question: "How does this assessment format (i.e., structured simulations) compare with others with which you have been evaluated (e.g., multiple-choice for CBEST and NTE Specialty Area Tests, classroom observation during students teaching) in terms of its assessment ability?" Roughly 50% (16 of 32) of those completing form A and 36% (12 of 33) of those completing form B gave answers that could be interpreted that they feel that the secondary general/life science assessment is better than the other assessments with which they have been assessed. Teachers particularly mentioned the CBEST and the NTE, the two multiple-choice assessments mentioned as examples. Sample comments are:

> *More valuable than multiple choice because it allows for more complete communication. Probably as good as classroom observation because one does not feel so much "on the spot" and having to play to an audience.*
>
> *This assessment is much closer to real-life examples of teaching. All of these tasks are what practicing teachers need to perform during the course of their job. Multiple choice tests are limited in that they only test to see if you can recognize the appropriate response.*
>
> *Structured simulation is a great idea, it really does test things a teacher does on an everyday basis. CBEST assesses our "professional skills" to see if we have minimum basic educated knowledge. NTE tests our knowledge of the content area -- "Biology"-- but I have yet to be tested for my ability to be "a teacher."*

Nine percent (3 of the 32) of the teachers completing form A and 6% (2 of the 33) of those completing form B believed that this assessment method was inferior, compared with the others. The following comments illustrate specific criticisms:

> *I feel the other tests (CBEST, NTE, etc.) were tests that could be scored fairly. These seem to be all gray areas.*
>
> *Believe NTE more fair and accurate. These assessments allow too many variables and ambiguities. Entirely too subjective.*
>
> *I feel the CBEST was a good indicator of basic skills, and I mean basic. I feel the NTE Specialty Test was a good indicator of Subject Area knowledge. I feel my student teaching evaluations were helpful in providing direction and in recognizing strengths and weaknesses. I feel that this test had nothing to do with the reality of teaching and would in no way be an indicator of my ability as a teacher.*

Twenty-two percent (7 of 32) of the teachers completing form A and 12% (4 of 33) of those completing form B did not offer an opinion of the relative merits of the Structured Simulation Tasks for Secondary Life/General Science Teachers and other teacher assessments. Instead, these comments indicated that the teachers believed that this assessment measured a different area than the specific assessments cited as examples:

> *CBEST and NTE test knowledge of subject not teaching skills. Class observation is similar to this as the observer is looking at how the person teaches and knowledge of subject. This test looks more at teaching skills than the other assessments.*
>
> *CBEST and NTE are more comprehensive evaluations of subject area knowledge, this is a better evaluation of classroom management and knowledge.*
>
> *One thing I like about these tests was that they were assessing teaching ability, not necessarily content knowledge such as in the CBEST and NTE. Knowing a subject does not mean that one can teach it.*

Considering summary responses to both the general question about the appropriateness of the assessment and the explicit comparison with other assessment methods, most teachers approved of this assessment. Again, the teachers completing form B tended to be more critical than the teachers completing form A.

Assessment Format

Format Features

This assessment format was a pencil-and-paper test with written stimuli which asked teachers to perform a series of tasks similar to those they encounter in their teaching. The tasks were developed through a "task shell" system where many different versions of a single task can be generated cost-effectively.

Clarity of Preparatory Materials

Prior to the assessment, teachers received information which gave them a limited idea of what they would be asked to do. Extensive preparatory materials were not developed for this pilot test. When teachers were contacted to solicit their participation in the assessment, they were told that the assessment consisted of approximately four hours of responding to structured simulation tasks with written prompts and written responses. The letter which they received confirming their participation described the assessment as consisting of "a set of structured simulation tasks depicting classroom management and instructional situations that a teacher of science might encounter." It also contained brief descriptions of six possible tasks (including one which was not pilot tested) and a list of 27 possible science topics that might be included.

Teachers were generally satisfied with the level of description provided in the preparatory materials (although this was probably significantly affected by the fact that their performance on the assessment had no consequences for them). As shown in Table 3.5, teachers generally believed that the preparatory materials were clear with respect to the description of assessment activities and the aspects of teaching being measured. They did not believe that the scoring criteria, which were not addressed at all in the preparatory information offered, were described clearly.

Teachers were also asked if there was any additional information that would have been helpful prior to the assessment. Eight teachers suggested that specific examples of the test items would be helpful, although some commented that it wasn't a problem (possibly

TABLE 3.5

TEACHER PERCEPTIONS OF THE PREPARATORY MATERIALS FOR THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

	Number and % of Teachers Responding that Various Aspects of the Assessment Were Described Clearly in the Preparatory Materials			
	Teachers Completing Task A		Teachers Completing Task B	
Aspect of Assessment Described	#	%	#	%
Assessment Activities	24	91%	29	88%
Scoring Criteria	15	47%	13	39%
Aspects of Teaching Being Measured	26	81%	24	73%
TOTAL N	32		33	

because the test had no consequences for them). Some teachers wished they had known other details such as whether the format of the test was written or oral and how and by whom it would be scored.

Clarity of Task Instructions

Because this assessment was in the developmental stage, the focus of the evaluation form was on identifying problems in the task instructions which could have affected teacher responses. Teachers were only asked to elaborate on their negative responses; few teachers elaborated on positive responses.

The clarity of instructions for each task was evaluated both through teacher reports and from observation of the scoring process for many of the tasks. Teachers were asked if the directions for each task were clear and, if they were not, to describe the difficulty experienced. As can be seen in Table 3.6, a majority of the teachers perceived the directions to be clear for each task, ranging from a low of 66% for **Parent/Student Letter** to a high of 88% for **Classroom and Facility Safety**. Each task will be discussed separately, combining teacher and scorer comments with FWL staff observations.

Teachers described several difficulties in completing **Applying Effective Instructional Techniques**, most of which could be addressed through revised instructions. Two teachers could not tell whether they were supposed to respond to everything the teacher said, or confine the response to what they considered significant. Another could not tell whether or not to limit the analysis to student/teacher interactions or whether instructional content should be critiqued as well. One teacher suggested that illustrating the method of recording responses for this task by a labeled example would reduce confusion.

Teachers sometimes made assumptions which were not warranted by the stimulus materials. Teachers were penalized for assumptions that contradicted the information provided. Responses that depended on assumptions that were consistent with the stimulus material but went well beyond the information provided were ignored. In some cases, scorers recommended modifying the script or providing additional contextual information to eliminate some possible interpretations of classroom events.

One aspect of **Applying Effective Instructional Techniques** which emphasized its artificiality was the "Dr. Jeckyll and Mr. Hyde" nature of the teacher whose behavior swung back and forth from being exemplary to extremely inappropriate. Some of the inappropriate

TABLE 3.6

TEACHER PERCEPTIONS OF THE CLARITY OF TASK INSTRUCTIONS FOR THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Task	Number and % of Teachers Indicating Directions to the Task Were Clear			
	Teachers Completing Task A		Teachers Completing Task B	
	#	%	#	%
Applying Effective Instructional Techniques	24	75%	23	70%
Teacher as Curriculum Decision-Maker	23	72%	-	-
Parent/Student Letter	21	66%	-	-
Lesson Planning	-	-	23	70%
Classroom and Facility Safety	-	-	29	88%
TOTAL N	32		33	

actions were also very obvious, such as when the teacher reprimands one student for applying makeup during class and ignores another.

For **Teacher as Curriculum Decision-Maker**, teachers reported difficulty in shuffling the many pieces of paper and figuring out how to complete the table in the answer sheet. The answer sheet was designed for ease in scoring, where teachers recorded code numbers for each activity. This made it difficult, however, to keep track of the activities already recorded without continually referring back to the papers with descriptions of activities.

For two administrations of this task, teachers were given scissors to cut apart the activities and physically reassemble them into the unit. Several teachers who experienced this version remarked on the amount of time it took to cut out the activities. The test developer had considered, and discarded, the idea of using index cards, but it is likely that this would solve some of the logistical problems that the teachers experienced.

One teacher suggested that Part II, the portion where the teacher provides a rationale for the activities selected, and Part I be completed simultaneously, as it was difficult to reconstruct the rationale after the fact. Another teacher did not understand the format in which responses were expected.

Both teachers and scorers reported that some of the estimated times for completing a laboratory were much too brief. The two scorers believed, in addition, that the descriptions of some of the lectures and films needed to be elaborated in order for the teachers to appropriately evaluate them.

The task which received the lowest percentage of teachers agreeing that the directions were clear was the **Parent/Student Letter**. Although one example was provided for each part, teachers reported being unclear on what was expected, requesting more examples. The scorers agreed that greater clarity in the directions as to the distinction between the two parts of the letter would have been beneficial, especially as it appeared that many teachers had never seen such a letter before.

Teachers also reported being unclear on whether the responses were to be in a list form or written out as it would appear in a letter.

Despite a list of the elements in the lesson to be covered, some teachers found the portion of the **Lesson Planning** task where they were to write their own lesson confusing. Some teachers wanted additional information, such as the length of the class period.

Another teacher wanted to know what kind of students composed the class; either they did not notice the brief description of the classroom or they wanted additional information.

Classroom and Facility Safety had the highest percentage of teachers reporting that the directions were clear; no teacher described any difficulties in completing this task.

Length of Tasks

Teachers were asked if they had sufficient time to complete each task, and to identify any task for which they needed more time. For Form A, only 28% (9 of 32) teachers reported sufficient time to complete all tasks. Twenty-four teachers suggested more time for **Teacher as Curriculum Decision-Maker**, six identified **Applying Effective Instructional Techniques**, and four singled out **Parent/Student Letter**. Two of these teachers reported needing more time to complete all the tasks. Estimates of the amount of additional time needed ranged from fifteen to thirty minutes. One dissenting teacher believed that the time limits should be reduced to forty-five minutes for each exercise.

Teachers completing Form B were more satisfied with the time allotted. Eighty-eight percent (29 of 33) of the teachers reported no difficulties in completing the tasks within the time limits provided. Four teachers identified **Applying Effective Instructional Techniques** as needing more time, and one chose **Lesson Planning**.

Clarity of Scoring Criteria and Procedures

Scorers were asked if they had any difficulties in applying the scoring criteria for any of the tasks. Only one specific problem was reported: "The biggest problem was knowing how much to 'read in' to answers (e.g., is 'denigrating primary language' the same as 'racial bias')." Two scorers noted that since the scoring guides were previously untried, many revisions were needed. To FWL staff who observed the scoring process, it seemed fairly straightforward to match teacher responses to specific scoring criteria; in some cases, it was difficult to apply the criteria which distinguished between the responses awarded one point and those awarded two points.

When scorers were asked if some tasks were harder to rate than others, only **Teacher as Curriculum Decision-Maker** and **Lesson Planning** were identified. However, each of these tasks was mentioned by both of the two scorers who graded them. **Teacher as Curriculum Decision-Maker** was described as "slow" or "tedious" to score. One part of that task, specifically referred to by one scorer, involved the application of a lengthy scoring algorithm

BEST COPY AVAILABLE

which checked for the presence or absence of numerous activities or patterns of activities in the unit plan. If the assessment were to be operationalized, that portion of the assessment would be keyed into a computer and scored through a scoring program. **Lesson Planning** was described as difficult to score because "candidates write poorly and have minimal skill in writing lesson plans."

Cost Analysis

Administration and Scoring Cost Estimate

The Structured Simulation Tasks for Secondary Life/General Science Teachers tasks are administered in a large group setting. Thus, the tasks can be administered by one or more persons with little or no training in the specific content of the assessment using procedures common to standardized group test administrations.

The largest component of the cost of this assessment is that of personnel. Scoring requires the training of raters knowledgeable in the content and criteria for the assessment. Scoring of the pilot test data, which included both training and actual scoring, required four days for two scorers for form A and roughly seven days for two scorers for form B. (For Form B, the scorers, who had also been part of the assessment development team, made extensive revisions in the stimulus materials and scoring criteria for some of the tasks. The tasks represented in Form B also had more subparts than those in form A.) We estimate this time as minimal to insufficient for training and scoring an assessment such as this. With more fully developed scoring criteria which can be extended to other tasks within the same task shell and more fully developed stimulus materials, it is likely that the system could be implemented on a wide scale basis. We will use the time and costs associated with scoring the pilot tests as the current best estimate for administering similar assessments.

The pilot test involved training four scorers and scoring 22-23 teacher responses (the remaining ten were used in training) to each of six tasks over a period of four to seven days. Training and scoring were conducted separately for each task. The amount of time required to score each task was more closely related to the number of its subparts than the length of time required by the teachers for its completion, but training, scoring, and some development work averaged 1 1/3 to 2 1/3 days per task, depending on the form. Based on this experience, we estimate that approximately two days per scorer would be required to train and score roughly 20 teacher responses to a single task. If a half day assessment consisted of three tasks, it would take approximately six scorer-days to score twenty teacher assessments. According to this logic, five scorers should be able to score 100 teacher assessments resembling either Form A or Form B in six days, with periodic checks to insure that scorers are applying scoring criteria correctly. Assuming a cost of $160 per day for each scorer, this implies a cost of approximately $48 per teacher to train scorers and score an assessment. If these same scorers were used again for a similar task shell, the training time might be shortened, reducing marginally the total scoring costs.

Costs for test administration, duplication of materials, postage, travel, etc. would also need to be added to the costs for scoring the assessments. As we have outlined on other assessments, a cost of $30 per assessment for these activities assume minimal travel costs for test administrators. A summary of cost estimates for administering and scoring an assessment like this include:

Training and Scoring:	$48 per assessment
Administration/Other:	30 per assessment
Total Administration and Scoring Costs:	$78 per assessment

Development and Pilot Testing Costs

The costs for developing the five tasks for this assessment were $130,157 and are broken out by cost category in Table 3.7, which also includes costs for pilot testing. These development costs are the expenses for the assessment developer to deliver prototype activities to the CTC and SDE. In addition, $45,211 was spent for the pilot testing of these tasks with 65 teachers.

These data provide a rough indication of the magnitude of costs that would be incurred if a similar assessment were to be adapted for implementation.

Technical Quality

This section describes the process by which the assessment was developed, and discusses the reliability and validity of the assessment based on analyses of teacher performance, and refers to other analyses which pertain to evidence of validity.

TABLE 3.7

DEVELOPMENTAL AND PILOT TEST COSTS FOR THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Cost Categories	Development	Pilot Testing
Staff-Salaries & Benefits	$54,202	$16,014
Consultants (Teachers, assessors, and other consultants)	0	2,853
Travel (Consultants and staff)	9,280	9,142
Other Direct Costs (Rand fee, site rental, phone, duplication)	27,650	7,731
Total Direct Costs	$91,132	$35,740
Indirect Costs	39,025	9,471
Total Costs	$130,157	$45,211

Development

The shell/task development process is iterative in nature. The process begins with brainstorming about the general features of a shell, but to facilitate mutual understanding of concepts, developers are encouraged to illustrate their ideas with concrete examples from their own teaching experiences. This requires relating broad generalizations about good and bad teaching practices to specific examples of it.

These discussions provide a bridge between general concepts about good teaching practice (the craft) and concrete teacher behavior. This bridge helps the team flesh out the essential elements of a particular task. The fleshing out process also identifies factors that need to be included in the shell, e.g., the generic types of stimuli to which the candidates should respond and which responses are more or less appropriate. The team's discussions therefore shift back and forth between a focus on the general features of the shell and the specific elements of a task that would simulate those features in a realistic way. Sometimes a task is developed before its shell because only through the task construction process can the elements be identified that need to go into the shell. Usually about 4 to 5 teachers (and teacher educators) participated in the task development process.

The assessment developers intended to pilot test the materials with at least six prospective or new teachers before the materials were released for larger scale pilot testing. However, in a few instances, due to circumstances beyond their control, the initial shakedowns failed to take place. These shakedowns are viewed as an integral part of the development process, which follows a cyclical model of develop, test, revise, test again until the prototype task is considered complete.

Reliability

The following analyses were performed on the pilot test data of 32 teachers for Form A and 33 teachers for Form B. Interrater agreements could not be computed for reasons explained below. Internal consistency estimates were generated to assess the degree to which the variables or factors within each of the tasks would form a measure and the degree to which the different activities related to each other and might form an overall assessment of a candidate.

Interrater agreements. The process followed in scoring was that the scorers conferred on instances where the scores differed by two or more points. Scorers then changed their original scores. Therefore, interrater reliability estimates were not calculated for the pilot test data since the ratings were not independent and reflected a consensus between scorers.

Internal consistency of the tasks and assessment. Coefficient Alpha reliability estimates were calculated for the tasks by using the individual ratings on subparts within each task. The reliabilities for the tasks and subparts are shown in Table 3.8. The reliability estimates for the tasks ranged from - .11 for **Teacher as Curriculum Decision-Maker** on Form A to .62 for **Lesson Planning** on Form B. These reflect a relatively low degree of internal consistency within the tasks. These results should be interpreted in light of the early and formative development of these measures. For example, the pilot test and scoring were used to further refine the stimulus materials and scoring criteria. A more positive interpretation of the low internal consistency is that the different subparts measure more independent factors of a teacher's performance. The lowest reliability estimate, for **Teacher as Curriculum Decision-Maker**, might be explained by the fact that it was clear that many teachers did not complete Part II, and 75% of the teachers identified this task as one needing more time for completion.

In judging the "goodness" of these data in light of the developmental status of the instrument, it is helpful to reflect that the developer built this prototype using experience and models used with other licensing examinations, particularly state bar examinations. The developer states, "On the surface, one would think that the scores on two tasks created from the same shell would correlate more highly with each other than would either of them correlate with the scores on tasks created from other shells. That may happen, but I doubt that the differences would be very large. The unique features of a task, such as grade level and subject matter for the unit, may be more familiar to some candidates than to others and thereby influence scores. For this reason and others, no one task, by itself, is likely to be very reliable. And, the correlation between tasks -- whether from the same or different shells -- will not be especially high (expect low .20's). Whether such a pattern of correlations is considered good or bad depends on the goals for the test. If all tasks correlate with each other to about the same degree regardless of whether or not they were created from the same shell, then this would undermine the position of those who want to use the test results for diagnostic and educational purposes, such as providing candidates with subscores for such things as 'lesson planning' or 'classroom management.' If, on the other hand, the purpose is to make a defensible pass/fail decision based on a general measure of teacher proficiency, then this pattern of intercorrelations is fine provided that as a group, the tasks simulate a wide range of important tasks that teachers should be able to perform and span the types of school contexts and subject matter areas to which the license applies."

TABLE 3.8

INTERNAL CONSISTENCY OF TASKS
SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Tasks	Task Reliability
Form A:	
Applying Effective Instructional Techniques	.50
Teacher as Curriculum Decision-Maker	-0.11
Parent/Student Letter	.23
Form B:	
Applying Effective Instructional Techniques	.55
Lesson Planning	.62
Classroom/Facility Safety	.30

Intercorrelations among tasks. Correlations among the three tasks of each Form were calculated for the 32 teachers completing Form A and 33 teachers completing Form B, and are reported in Table 3.9. Only the correlation between the **Teacher as Curriculum Decision-Maker** and **Parent/Student Letter** was statistically significant. This pattern is again consistent with what the developer had predicted. If this pattern were to persist with further development and refinement of the assessment, it would imply that an overall decision using information across tasks would be based on multiple, relatively independent factors rather than an overall composite measure of a teacher's ability. As stated earlier, either pattern, i.e., multiple or single factors, is acceptable but the type of information and its use should be interpreted in light of the pattern(s).

Validity of Agreement Through Group Comparisons

Teachers participating in the pilot test represented different ethnicities, gender, teaching experience, etc. Examining differences among these might provide some tentative information about the validity of the assessment. For example, positive evidence would include that differences among ethnic or gender groups are minimal and differences among teachers with more or less experience and preparation support the assessment's sensitiveness and ability to measure any additional knowledge the training and experience might provide. Although differences between groups would be difficult to detect given the relatively low reliabilities associated with the current assessment, it may still be worthwhile to examine the differences for any patterns. Table 3.10 contains a summary of the trends for the pilot sample of 65 teacher candidates. Appendix A provides the means, standard deviations and numbers of candidates from which these summaries were constructed. A plus (+) indicates that the mean or average for the first group was greater than that for the second group. For example, the pluses under the Female-Male column indicate that for 3 of 6 tasks, the average female score was greater than that of the males. No notable differences were detectable on any of the variables where the groups were evenly split (3-3) or nearly evenly split (2-4). Whether the lack of differences is due to the characteristics and status of the assessment or due to the absence of differences among the groups is unknown at this point.

With further development, it would be desirable to observe patterns such that teachers with more training and experience outperform those with less and that scores of teachers of different gender, ethnicity, or teaching location are not notably different.

Content validity. Evidence of the content validity of this assessment comes from three sources. The first is the role that teachers and science educators have had in its

TABLE 3.9

INTERCORRELATIONS AMONG TASKS
SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Tasks	Correlations I	II	III
Form A:			
Applying Effective Instructional Techniques	--		
Teacher as Curriculum Decision-Maker	-.04	--	
Parent/Student Letter	.31	.39*	--
Form B:			
Applying Effective Instructional Techniques	--		
Lesson Planning	.11	--	
Classroom/Facility Safety	.11	.09	--

*$p < .05$

TABLE 3.10

TRENDS OF MEAN DIFFERENCES IN TASK PERFORMANCE BETWEEN CANDIDATES WITH DIFFERENT CHARACTERISTICS*

SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

Activity	Gender Female/ Male	Teacher Preparation Regular/ Intern	Level of Teaching HS/Middle or Jr.High	Teaching Location Inner-City Other	Ethnicity Non-Minority/ Minority
Form A					
Applying Effective Instructional Techniques	-	+	+	-	-
Teacher as Curriculum Decision-Maker	+	-	-	+	-
Parent/Student Letter	+	-	-	+	-
Form B					
Applying Effective Instructional Techniques	+	+	+	-	+
Lesson Planning	-	tie	-	-	+
Classroom and Facility Safety	-	+	+	-	-
SUMMARY	3/6	3/6	3/6	2/6	2/6

*Entries reflect the direction of the mean differences for the different candidates. For example, for Applying Effective Instructional Techniques, Form A, the average mean of male teachers in the pilot test was greater than the females. The individual differences for each task or activity do not generally represent statistically significant changes.

development. The second is the analyses of the match of the assessment to the model curriculum guide and California Beginning Teacher Standards that compares the assessment's content with that recommended in the official documents. The third is the type of concerns raised by the beginning teachers who participated in the pilot test. These analyses have been described earlier and implications for further development are described in the following section.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Structured Simulation Tasks for Secondary Life/General Science Teachers, organized into the areas of administration, scoring, content, format, and a brief summary.

Administration of Assessment

Like other large-scale examinations, the Structured Simulation Tasks for Secondary Life/General Science Teachers is administered simultaneously to a large number of people. Benefitting from many years' experience in conducting such examinations, the administration of the actual assessment poses few logistical problems. The only difference between this assessment and traditional large-scale tests is the requirement of additional surface space to accommodate the materials for each task. Although no trouble was experienced in locating facilities for this small pilot test, the additional space requirement may preclude the use of large lecture rooms or auditoriums equipped with small, easily-stored writing surfaces.

Our experience in locating secondary life/general science teachers to participate in the assessment leads us to conclude that such teachers do not tend to be concentrated in concise geographic areas, even within large metropolitan areas. The administrative requirements of the assessment make it possible to be centrally administered to large groups, thus considerably reducing the administrative costs per teacher. However, it should be noted that the higher degree of centralization afforded by this assessment may place larger burdens on teachers from rural areas and the outer edge of metropolitan areas who would have to travel a long distance to a selected site.

The development of a "shell" for each task permits the teaching context and science content to be varied while ensuring the comparability of tasks across time. Tasks where the correct answers are relatively independent of the content and context portrayed would pose security risks over time due to their use of formulaic answers which could be easily memorized. Tasks whose answers are highly dependent on the content and context portrayed are more suitable for variation over time.

Scoring

Scoring consists of checking teacher responses against a predetermined list of possible correct responses. Scorers judge responses which are not on the list according to their professional judgement, and are free to award credit to responses judged to be acceptable which are not on the original list. For the most part, this methodology worked well, although major revisions in either the scoring criteria or stimulus materials were needed for some of the tasks. This suggests the need for more extensive pilot testing of tasks prior to their administration as an assessment.

Scoring training consisted of an orientation to the scoring guide, independent scoring of a sample response, and a group discussion of the resulting scores. Approximately ten sample responses were scored for each task subpart. This represents a departure from the assessment developer's recommended practice, where approximately fifty sample responses are used in the training.

FWL has the following recommendations for revisions in the scoring training:

- Provide more examples of scored responses, especially for those task subparts where partial credit is given for incomplete responses.
- Although the scorers believed that minimal knowledge of science and science teaching is necessary to score the tasks, scorers should continue to be recruited from experienced science teachers until data is available to consider the effects of the use of scorers with lesser qualifications.

If the ultimate set of tasks which constitute the assessment represent a sufficiently broad sample of tasks that are critical to teaching success, the Structured Simulation Tasks for Secondary Life/General Science Teachers should be sufficient for purposes of licensure. However, since there is little information on specific teaching competencies either within or across tasks, this assessment is less useful for yielding diagnostic information for staff development or beginning teacher support.

Assessment Content

Our observations and information collected from scorers and teachers participating in the pilot test suggest the following conclusions about content:

- Modifications of the tasks are necessary to bring the assessment in to closer congruence with the latest Science Framework, especially with respect to the lack of variety in the science content portrayed, no representation of thematic structuring of science content, and the lack of representation of the middle school curriculum.

- Coverage of the California Standards for Beginning Teachers varies. Every standard is addressed to some extent. The standards addressing curricular and instructional planning skills, student diagnosis, achievement and evaluation, and cognitive outcomes of teaching are most completely addressed. Standards receiving limited attention include student motivation, involvement and conduct, affective outcomes of teaching, the capacity to teach crossculturally, and professional obligations.

- Most of the teachers and all of the scorers believed that the tasks were relevant to the job of secondary science teachers.

- Most of the teachers agreed that they had been sufficiently prepared to respond reasonably to the tasks. However, over one-third of the teachers reported difficulty with one or more of the tasks. The largest number of teachers identifying specific tasks as difficult cited the two tasks related to instruction, **Lesson Planning** and **Teacher as Curriculum Decision-Maker**. Scorers identified **Lesson Planning**, **Classroom and Facility Safety**, and the **Parent/Student Letter** as being particularly difficult for the teachers participating in the pilot test.

- Teachers held mixed opinions as to whether or not the assessment was appropriate for teachers of different grade levels. Concern was expressed for middle school or junior high school teachers, since the curriculum at that level was not represented in the assessment.

- Between half and two-thirds of the teachers believed that the assessment was appropriate for teachers of diverse student groups. Teachers' criticism of the assessment reflected discomfort with being asked to design instruction for students they had never taught or to evaluate teaching methods which they did not use.

- An expert on teaching diverse students cautioned that unless culturally diverse models of appropriate teaching are built into the scoring criteria, scorers may be unable to recognize culturally appropriate responses that lie outside their range of cultural experience and/or knowledge. Teachers from particular cultural communities who are teaching effectively in that community may be penalized as a result.

- Teachers and scorers overwhelmingly believed that the assessment was fair to different groups of teachers. The expert on teaching diverse students, however, cautioned that unless every effort is made to reduce possible instances of miscommunication, teachers who are not from the dominant culture may not display the full extent of their knowledge about teaching.

- Between one-half and three-quarters of the teachers believed that this assessment method is appropriate for measuring their skills as secondary science teachers.

Assessment Format

The assessment format is a pencil-and-paper test with written stimuli which asks teachers to perform a series of tasks similar to those encountered in teaching; responses are compared to a predetermined set of correct responses.

Based on evaluations by teachers, scorers, and FWL staff, the following modifications in the tasks are needed:

- Directions for **Applying Effective Instructional Techniques** should be revised to more clearly indicate the type and form of responses teachers should make. The teacher's behavior in the script should be more plausible, i.e., not shift from exemplary to extremely inappropriate. Roughly 15% of the teachers identified this task as needing more time for completion than the hour provided.

- For **Teacher as Curriculum Decision-Maker**, teachers should be given the activities printed on index cards to facilitate sorting. Descriptions of activities and estimated times for their completion should be double-checked; descriptions should contain enough information to enable the teacher to understand their content and possible strengths and weaknesses. Three-fourths of the teachers reported needing time beyond the ninety minutes provided to complete this task.

- Directions and examples which make a more clear distinction between the responses expected in the two parts of the **Parent/Student Letter** are needed.

- Although some teachers believed that additional information was needed in order to complete **Lesson Planning**, almost all of the additional information cited was already provided. Some experimentation with the format in which it is presented may be needed in order to assist teachers in locating it.

- No teachers reported difficulty in completing the **Classroom and Facility Safety** task.

Summary

The methodology used in the Structured Simulation Tasks for Secondary Life/General Science Teachers has been successfully implemented in the application portion of examinations for licensure of lawyers. While minor revisions are still needed to obtain a fully developed prototype, pilot test results suggest that it could be successfully replicated in teaching. However, the nature of its scoring system, while suitable for licensure decisions, is less suited to yielding diagnostic information to inform staff development and/or beginning teacher support.

CHAPTER 4:

SCIENCE LABORATORY ASSESSMENT

The Science Laboratory Assessment is an observation system developed by RMC Research Corporation in Mountain View, California. As its name suggests, the context of the assessment is a science laboratory activity. This activity may be conducted either in or outside the classroom (e.g., a science field trip), but it must be student-centered, hands-on, and inquiry-oriented. The major portion of the assessment consists of an observer using the Science Laboratory Assessment instrument to conduct a 30-45 minute (minimum) observation of the science laboratory activity, focusing on seven domains of teacher performance.

The seven domains of teacher performance are deliberately broad in scope to represent aspects of teaching at all grade levels, in all subject areas, and in a variety of settings. The seven domains are: **Pedagogy**, **Content**, **Materials/Equipment**, **Management**, **Knowledge of Students**, **Climate**, and **Communication**.

For each of the seven domains, there are from between two and nine elements which help define the domain being observed. An example of one domain and its four defining elements is as follows:

Domain:	**Materials/Equipment**
Elements:	*Teacher Use, Safe Setup, Safe Practices, Availability*

Each element is further defined by indicators which describe the type of teacher performance to be observed and examples of behaviors or events that provide evidence for use in the assessment. A complete list of the assessment's domains and elements can be found on Figure 4.1, page 4.2. Figure 4.2 gives an example of the defining indicators for three elements of the **Materials/Equipment** domain. (See Appendix B for a complete description of the domains, elements, and indicators, as well as the materials used for the documentation and analysis process.)

FIGURE 4.1

LIST OF DOMAINS AND ELEMENTS: SCIENCE LABORATORY ASSESSMENT

A. PEDAGOGY

- A1. Planning
- A2. Sequence
- A3. Prelab
- A4. Directions
- A5. Explanation/ Presentation
- A6. Monitoring/Adjusting
- A7. Feedback
- A8. Questioning
- A9. Closure

B. CONTENT

- B1. Accurate
- B2. Integrated
- B3. Related to Objectives

C. MATERIALS/EQUIPMENT

- C1. Teacher Use
- C2. Safe Setup
- C3. Safe Practices
- C4. Availability

D. MANAGEMENT

- D1. Grouping
- D2. Other Personnel
- D3. Routines and Transitions
- D4. Student Engagement
- D5. Timing
- D6. Student Behavior
- D7. Lab Cleanup

E. KNOWLEDGE OF STUDENTS

- E1. Diversity
- E2. Student Characteristics

F. CLIMATE

- F1. Interactions with Students
- F2. Interactions among Students
- F3. Attitudes
- F4. Inquiry

G. COMMUNICATION

- G1. Speaking
- G2. Writing
- G3. Listening
- G4. Strength of Presence

FIGURE 4.2

THREE ELEMENTS AND DEFINING INDICATORS OF THE MATERIALS/EQUIPMENT DOMAIN: SCIENCE LABORATORY ASSESSMENT

Elements	Defining Indicators
C1. TEACHER USE	The teacher properly uses the equipment and handles the materials employed in the observed laboratory activity. Live organisms are maintained and handled in a humane and appropriate manner. Where applicable, the teacher is alert to student allergies, fears, and other problems related to the use of specimens or live organisms in the science lab activity.
C2. SAFE SETUP	The setup of equipment, furniture, and materials has no serious irregularities or dangerous conditions. The setting has, as needed, adequate ventilation, first aid supplies, safety equipment, corrosive-resistant counter tops, a fire extinguisher, running water, good lighting, etc. Materials and equipment are stored, labeled, and moved properly.
C3. SAFE PRACTICES	The teacher knows about the potential dangers involved in the planned science laboratory activity. The teacher informs students about, checks for understanding of, and enforces the proper use of equipment and handling of materials, as needed. The teacher tells students about safety procedures, potential dangers and actions to take, and proper cleanup and disposal procedures. Students are wearing safety gear (e.g., goggles, aprons, gloves) when needed. Cleanup and disposal are completed in a well-coordinated and safe manner. The teacher is alert to potential safety problems, knows what to do if a safety problem occurs, and takes corrective measures when necessary. There are no observed teacher violations of state and federal safety laws and regulations on the setup, use, and handling of materials and equipment.

Although the majority of evidence corresponding to the domains, elements and indicators comes from the actual observation, the Science Laboratory Assessment also provides for evidence to be collected by the observer from three other assessment components: (a) a Pre-Observation Questionnaire completed by the teacher, (b) a 20-30 minute Pre-Observation Conference with the teacher, and (c) a 15-20 minute Post-Observation Conference with the teacher.

As is often the case with high-inference, observation instruments, a key feature of the Science Laboratory Assessment is its documentation and analysis process. This process entails extensive scripting during the observation and then a rewriting of the data in a specific manner on another form. These two steps must be done before the observer gives any ratings of the teacher's performance.

A distinctive feature of the Science Laboratory Assessment is the part of its documentation process called guided note-taking. Instead of requiring the observer to script the entire lesson as accurately as possible in a chronological manner (as is done with some high-inference observation instruments), the guided note-taking process requires the observer to categorize the evidence and notes from the lesson at the same time as it is scripted. That is, the observer categorizes the information from the observation by domain simultaneously with recording it. To facilitate this procedure, the observer scripts all evidence and notes on a specially-designed Guided Note-taking Form (GNF) which is divided into seven spaces corresponding to the assessment's seven domains (see Appendix B). Typically, an observer may use 12-15 of these forms to record data from a single observation.

Other distinctive features of the Science Laboratory Assessment are its Documentation Sorting Record and a Summary Report Form. Upon completion of the observation, the observer takes the information from the Guided Note-taking Forms and further categorizes (or sorts) the data by element on a seven-page Documentation Sorting Record (DSR). Data from the pre- and post-observation conferences and the questionnaire is also sorted by element on the DSR.

After reviewing all the information on the DSR, the observer uses a Summary Report Form to rate the teacher's performance on each domain and to enter an overall rating.

Using a two-point rating system, the observer is instructed by the Assessor's Handbook to give a rating of "2" if "s/he feels the teacher's performance is minimally acceptable or better," and a rating of "1" if the performance is not acceptable. For each rating, the observer is also asked to provide three or four corresponding summary remarks.

For this pilot test, a single observation per teacher was deemed sufficient for trying out the assessment instrument. If this assessment were to be used as the primary data for credentialing purposes, the developers of the assessment suggest that a minimum of four observations should be conducted for each new teacher, with the new teacher conducting a different type of lab activity (e.g., exploratory, illustrative) in different content areas each time.

The administration of the Science Laboratory Assessment in this pilot test, the content of the instrument, and the assessment format are discussed below. The content and format sections of the report contain information from the teacher and assessor evaluation forms, as well as information and analysis of scoring results. Following these are discussions on cost analysis and technical quality of the prototype assessment. The chapter concludes with an overall summary with recommendations for further steps in exploring the feasibility and utility of high-inference, subject-specific observation instruments such as this in California.

Administration of Assessment

Following an overview of the administration of the assessment, this section contains information on the following: logistics (e.g., identifying the teacher sample, scheduling observations), security, assessors and their training, scoring, and perceptions of the administration by teachers, assessors and FWL staff members.

Overview

As with any observation system, the administration of the Science Laboratory Assessment required careful planning and coordination on the part of the observers, the new teachers, and the school administrators. Observers and new teachers had to be recruited and scheduled, and observers also had to be trained. Moreover, because this observation system is content-specific, new teachers had to be carefully matched with assessors of the appropriate science background (e.g., life or physical science). Over approximately a six-week period which ended in June, 1990, eleven trained assessors observed a total of 29 new teachers conducting laboratory science lessons. As shown in Table 4.1, the observations were conducted in five areas across the state, and both Project and Non-Project teachers participated. Although the majority of the 29 participating teachers were white, in their first year of teaching, and teaching at the secondary school level (i.e., high school or middle school), the teacher sample was almost evenly divided among males and females.

Logistics

Administration of the Science Laboratory Assessment required the following logistical activities: recruiting and training observers, identifying the teacher sample, scheduling the observations, sending orientation materials to the teachers, and acquiring evaluation feedback from the teachers and assessors.

Recruiting and training observers. The observers for the RMC pilot test were carefully recruited by FWL staff with assistance from a consultant to the Science Laboratory Assessment project. All were experienced science teachers, either currently teaching at the secondary level (i.e.. middle school or senior high), working at the district level, or on sabbatical. Twelve observers were originally recruited, three of whom were members of the Science Laboratory Assessment Development Committee. Unfortunately, one of these three had to withdraw from the program after the training due to time constraints. RMC staff trained the observers in a two-day session, preceded by one day of home study. (For more information about the observers and their training, see the section, "Assessors and Their Training.")

Identifying the teacher sample. Table 4.1 presents information about the teacher sample for this assessment. It was necessary to recruit Non-Project teachers in addition to CNTP teachers in order to have a sample that represented different regions of the state, all grade levels (i.e., high school, middle school and elementary), different ethnic groups, and a variety of teaching contexts (including physical or life science classes). As was the case last year, FWL staff recruited the majority of the Non-Project teachers by calling school districts which neighbored CNTP districts, and asking for names of first- and second-year science teachers.

TABLE 4.1

PILOT TEST PARTICIPANTS
SCIENCE LABORATORY ASSESSMENT
(Number of Teachers = 29)

Location	No. of Teachers		Teacher Characteristics
	Project	Non-Project	
Chico Area	2	-	24 Caucasian, non-Hispanic; 2 Hispanic; 1 Asian or Pacific Islander; 1 Native American; 1 Other 16 Male; 13 Female 13 High School; 11 Middle School; 5 Elementary School 17 First-Year; 10 Second-Year; 2 Third Year
Sacramento Area	1	3	
San Francisco Area	5	3	
Fresno Area	3	-	
Los Angeles Area	1	3	
Total Number of Teachers	12	17	

Scheduling the observations. After identifying the participants, the observations were scheduled. Scheduling required a match between the teacher and observer in three areas:

(1) availability (e.g., dates, times);
(2) science background (e.g., life or physical science); and
(3) teaching background (i.e., secondary or elementary teaching experience).

A teacher teaching a high school chemistry lesson, for example, had to be paired with an observer who also had high school chemistry experience and who was available to observe on the date and at the time the teacher chose. Similarly, an elementary teacher conducting a life science lesson could only be paired with an available observer with a life science background and who also had some elementary school experience (i.e. teaching elementary school or elementary teachers, or developing elementary science curriculum). Moreover, in order to get a double-scoring sample, five teachers were observed by five different pairs of observers. For these observations the logistical difficulties were increased.

Sending orientation materials. Shortly before the observations, the participating teachers were mailed an orientation handbook which included the domains, elements, and indicators of the assessment, the questions for the Pre- and Post-Observation Conferences, and the Pre-Observation Questionnaire to be filled out by the teacher before the observation. Also included were three samples of completed Questionnaires, each representing a different grade level (i.e., elementary, middle, and high school).

Collecting evaluation feedback. After the observations, teachers were sent an evaluation form to fill out and return to FWL. Evaluation forms were also given to each of the observers who returned them to FWL along with their observation documentation.

Security

Because the content of the assessment was included as part of the orientation materials, the focus of security for this assessment was on the completed documentation for each teacher. Assessors mailed the documentation materials to FWL, where they were securely filed.

As we noted in the first-year report (Assessment Component of the California New Teacher Project: First Year Report, March 1990), if an observation system is selected as a method of assessment for credentialing teachers in California, procedures to ensure security at the observation and processing stages (and during long-term storage) would have to be developed and implemented. Each piece of documentation (i.e., Guided Note-Taking Forms, Document Sorting Records, Summary Response Forms, Pre- and Post-Observation forms, and Pre-Observation Questionnaires) would have to contain identifying information in case the pieces became separated. For this type of assessment, such information would probably include the following: teacher code, observer code, date of observation, and perhaps type of science lesson. All documentation for a given teacher credential candidate would also have to be retained for a minimum number of years, enough to cover the period in which teachers could appeal decisions, or to meet statutory requirements.

Assessors and Their Training

Twelve assessors were recruited and trained to conduct and score a minimum of three observations each for this assessment. As mentioned earlier, one of the assessors withdrew from the program after the training due to time constraints. This section describes some characteristics of the assessors, describes the training, and presents the perceptions of the training by the assessors and FWL staff.

Characteristics of the assessors. All of the assessors had several years experience as a teacher in California schools, were knowledgeable about at least one area of science, and had worked with student teachers, as a teacher trainer, or as a mentor teacher. Of the eleven observers who participated in the pilot test, there were five females and six males. Nine of the observers were currently teaching high school science; of the three other observers, one taught junior high science (and served as department chair), one was a district science resource specialist, and the third was on sabbatical working at a private chemical laboratory. All but two of the observers worked in Northern California; one worked in Fresno and the other in Southern California.

Training. Training for conducting and scoring the observations lasted three days: one day of home study, and two days of group training. For the home study day, trainees were instructed to read the "Observer's Handbook," and to thoroughly familiarize themselves with the domains, elements and indicators, and with the assessment materials, forms, and procedures.

The two-day group training was conducted by RMC staff in San Francisco on April 6 and 7, 1990. The first day of training consisted of the following: a quick review of all the documentation forms (e.g., Pre- and Post-Observation Conference Forms, Guided Note-Taking Forms); training in how to conduct the Pre-Observation Conference, followed by role playing; and practice in using the Guided Note-Taking Forms by watching and taking notes from videotape segments of science lessons.

The second day of training provided the trainees with more practice using the Guided Note-Taking forms (using videotape segments); practice in completing a Documentation Sorting Record; an introduction to the Summary Report Form (i.e., how to score performances), and simulated practice (again using videotapes) in conducting, documenting, and scoring an observation.

At the end of the two days, the trainees' documentation from the final practice videotape was collected for review by the trainers. This documentation was used by the trainers and FWL staff to informally assess each trainee in three areas: (1) recording evidence and notes properly on the Guided Note-Taking Forms; (2) sorting the information by elements in a reasonable manner on the Documentation Sorting Record; and (3) making objective remarks and reasonable judgements on the Summary Report Form. Based on the problems trainees encountered, RMC and FWL staffs prepared a three-page list of "helpful hints" which was mailed to all observers before they did any observations for the pilot test.

Perceptions of training. The observers were not asked by the trainers to evaluate their training at the end of the session. When collecting evaluation feedback from the observers, however, FWL staff included a page of questions about the training they had received. Of the eleven observers, one observer said the training as "very good," six observers described the training as "adequate," and three found the training to be "insufficient." Six of the observers also noted that they found the practice in taking notes from the videotapes to be the most useful part of the training.

All of the observers had suggestions for improving the training. Suggestions for improvement given by more than two observers were as follows:

- Increase training time (5 observers)
- Increase opportunities for discussion about the assessment (4 observers)
- Make sure the assessors understand the domains and elements at the beginning of the training (2 observers)

Based on their own observations of the training, FWL staff concur with the above suggestions. Training would be improved by increasing the training time, especially the time allotted for "hands-on" practice (e.g., practice using the different forms, practice scoring the candidates) and discussion of the assessment's different components (e.g., documentation, scoring). In addition, any future training for this assessment should begin with a review of, or solid introduction to, the assessment's content so that all of the participants agree on the definitions of the elements. Since the content is the foundation of the assessment, this review should be done before giving the observers any other task (e.g., asking them to watch videotape segments and collect evidence pertaining to the elements).

Training could also benefit from more explicit instruction and examples on how to record evidence and notes on each of the assessment forms, as well as on how to evaluate teacher performances. Regarding the latter, observers should also be given ample opportunity to practice and discuss the evaluation process in order to help ensure that there is consistency among observers. (More discussion of the evaluation process can be found in the next section, "Scoring.")

In order to provide time for the improvements described above, the two-day group training should be extended by a minimum of one full day. As evidenced by the amount of questions, confusion, and actual frustration expressed on the first day of the group training, the home-study day seemed to contribute little to the training, aside from a basic introduction to the assessment's content, forms, and procedures. Although a home-study day was chosen by the trainers because it was very difficult to schedule three consecutive days for training which all the observers could attend, at least three consecutive in-session days of training appear necessary if the administration, content, and format of the assessment are to be covered adequately.

Scoring

The scoring system of the Science Laboratory Assessment is an integral part of the assessment prototype. The same person who conducts the observation uses the documentation from the observation to score the assessment.

For the Science Laboratory Assessment, the scoring process is also directly linked to the documentation process. The observer first records data (i.e., evidence and notes) from the observation directly on the Guided Note-Taking Forms. These forms, as mentioned earlier, require the observer to categorize the data according to domain. Upon completion of the observation, the observer begins step two which requires further categorization of the data by "sorting" it on the Documentation Sorting Record (DSR). The observer sorts the data according to the elements which correspond to each domain. The observer also uses the DSR to categorize and sort by element the data collected on the teacher's questionnaire and on the two conference forms.

Having sifted or sorted the data according to the elements corresponding to each domain, the observer is then ready to make judgments about the teacher's performance level in each domain. Taking one domain at a time, the observer first reviews all the information on the DSR which corresponds to that domain. If, looking at all the information listed across the elements of the domain, the observer "feels the teacher has shown a minimally acceptable level of performance," then he observer writes a "2" in the rating box for that domain on the Summary Report Form. If the observer feels the teacher's performance is not acceptable, the observer enters a "1." If the observer can not make a judgement, either because of lack of information or a borderline level of performance between minimally acceptable and not acceptable, the observer enters an "X." After giving a rating, the observer writes three or four summary remarks about the teacher's noteworthy strengths and weaknesses related to the elements in that domain. The observer repeats this process for all seven domains.

The last step of the scoring process requires the observer to make an overall judgment of the teacher's performance. The observer again rereads the information on the DSR, and then reviews his/her judgments and remarks made on the Summary Report Form. After reviewing all of this information, the observer makes a judgment as to whether the teacher's overall performance is acceptable (a "2" rating) or unacceptable (a "1" rating). Any comments the observer thinks should be considered regarding the overall rating are entered next to the rating.

The scoring process, like those of other high-inference observation systems, is very labor intensive. Not including the observation time, the entire process takes, on average, approximately three to four hours.

Teacher, Assessor, and FWL Staff Perceptions of Administration

All but one of the teachers and one of the observers expressed satisfaction with the arrangements (e.g., scheduling, room arrangements) made for the administration of this assessment. The one dissenting teacher did not like being assessed at the end of the school year. The one dissenting observer--an observer who conducted four observations for the pilot test--stated that she had "absolutely no complaints with the logistical arrangements," but that she did not like having to leave her classroom in order to administer the assessment. As she explained,

> *I had difficulty squeezing out the time to make the carefully documented lesson plans my substitutes needed so that I could be away [to observe]. In addition....I had to expend considerable effort in advance planning so that I could create the kind of day a sub could handle.*

Concern, similar to the above assessor's, about trying to juggle the administration of these assessments with the execution of their teaching duties was also expressed verbally to FWL staff by other observers. Although the assessors' burdens would be lessened if the observations were spaced further apart (all observations for this pilot test were conducted within a six-week period during a busy time of the school year), good teachers devoted to their students would probably still experience difficulty leaving their classrooms, especially if they were also taking time off for other professional obligations (e.g., serving as mentor teachers, serving on curriculum development committees).

Should an observation system such as this be considered for credentialing use in the state, the time difficulties experienced by the observers in this pilot test should be kept in mind. The issue is especially noteworthy if one agrees with the assessors of this pilot test, all of whom stated that this assessment should only be administered by experienced science teachers. Although the assessors differed as to how much experience is necessary (answers ranged from "moderate" to "a great deal"), all agreed that the assessment should not be administered by someone who is knowledgeable about science but has no science teaching background. As one observer noted, "Science trained non-educators haven't a clue about classroom management and planning."

Finally, with regards to administering the assessment, nine of the eleven assessors expressed displeasure with the amount of time it took to complete the Documentation Sorting Record, and eight assessors said they had difficulty observing the lesson and collecting evidence. Since both of these issues are directly related to the format of the assessment, they will be discussed more completely in the section, "Assessment Format."

Assessment Content

The developers of the Science Laboratory Assessment chose science, and in particular, science laboratory activity, as the content and focus of this assessment for several reasons. First, responding to the national need for highly trained scientists, they believe that in order to develop top scientists, we need top quality science teachers, starting at the elementary school level. Second, research conducted by two of the assessment developers reveals the importance of science instruction for developing students' basic skills in reading, mathematics, and writing (Wheeler, 1986-87), as well as more advanced thinking skills in these and other areas (Quellmalz, 1985). Third, educators who were asked by the California Commission on Teacher Credentialing to review the two NTE Specialty Area Tests in science (i.e., Biology and General Science; Chemistry, Physics, and General Science) expressed great concern that, in the credentialing process, there was no assessment of laboratory demonstration and presentation skills or of lab safety. While acknowledging that some aspects of lab safety could be assessed by a multiple-choice exam, all of the reviewers felt that the only way to evaluate the laboratory presentation and demonstration skills of credential candidates was through a performance assessment which focused on laboratory science.

In the following pages, the content of the Science Laboratory Assessment is evaluated along these dimensions:

- Congruence with the 1990 California Science Framework;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the instrument;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, diverse student groups);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

We would like to note that, as was the case with all of the assessment instruments pilot tested this spring and summer, the Science Laboratory Assessment was developed for the State of California within a specific development timeline. Although the science educators who participated in the statewide review of the instrument were asked to comment on the job necessity and appropriateness for new teachers of the domains and elements covered by the assessment, there was not sufficient time to conduct a larger content validity study. Without such a study, our ability to comment on the assessment's appropriateness along such dimensions as job-relatedness, appropriateness for beginning teachers, and appropriateness across contexts is limited. Thus, excluding the first two dimensions of curriculum congruence and standards coverage (which are based on FWL staff's analysis of the documents involved), the discussions of the remaining dimensions are based on the perspective of the participating teachers and assessors, and FWL staff, as reflected in feedback forms, in informal conversations with the assessors, and in data from the assessment's rating forms.

Congruence with the 1990 California Science Framework

FWL staff reviewed the Science Laboratory Assessment to see in what ways the assessment is congruent with the *California Science Framework*, and how it could be modified to achieve better congruence. For our analysis, we used the *1990 Science Framework for Kindergarten through Grade Twelve*. This framework is divided into three parts, each focusing on a different aspect of science instruction. The first part provides a context for instruction by describing the nature of science and the major themes of science. The second part focuses on instructional content, providing examples of theories and themes in the life, physical, and earth sciences to be taught at different grade levels (K-12). The third and final part of the framework presents specific information on how to achieve a desired science curriculum. It includes descriptions of appropriate science pedagogy to be applied by the teacher in the classroom; ways in which a district or school can implement a strong science program, and the criteria used by the state in its adoption process of science instructional materials.

Table 4.2 lists the three parts of the framework and their corresponding chapters, and then describes the Science Laboratory Assessment components (e.g., domains and elements, pre-observation conference questions) that are congruent with the framework. As the table indicates, there are some assessment components that are congruent with each part of the framework, but only in a partial manner. Strongest congruency is with the framework's description of science processes and the teaching of science (Chapter 6). This part of the framework is addressed by the nature of the assessment (i.e., an observation of a hands-on science laboratory activity), several domains and elements, and some questions on the pre-observation conference form. Even this congruency is partial, however. One part of the chapter, for example, presents those science processes (e.g., observing, communicating,

TABLE 4.2

COVERAGE OF THE CALIFORNIA SCIENCE FRAMEWORK BY THE SCIENCE LABORATORY ASSESSMENT

Content	Relevant Assessment Components
PART I: WHAT IS SCIENCE?	
Chapter 1: Nature of Science	-Addressed by the "Attitudes" element of the Climate domain.
Chapter 2: Major Themes of Science	-The "Integrated" element of the Content domain requires that the teacher knows the theme of the activity. #14 on the Teacher Questionnaire asks the teacher to specify the theme of the activity.
PART II: CONTENT OF SCIENCE	
Chapter 3: Physical Science Chapter 4: Earth Sciences Chapter 5: Life Sciences	-The content of the observed activity should fit into one of these three areas.
PART III: ACHIEVING THE DESIRED CURRICULUM	-Chapter 6 content addressed by several domains and elements: Pedagogy ("Planning," "Questioning"); Knowledge of Students ("Diversity," "Student Characteristics") and Climate ("Interaction with Students", "Inquiry")
Chapter 6: Science Processes and the Teaching of Science	-Also addressed by Pre-Observation Conference Questions #5, 6, and 10. -Observation must be of hands-on activity.
Chapter 7: Implementing a Strong Science Program (at school district and site level)	-Not addressed by this assessment.
Chapter 8: Instructional Materials Criteria (as applied to adoption of materials)	-Not addressed by this assessment.

comparing) which can best be expected from students at different grade levels. Although the Science Laboratory Assessment includes references to science processes in at least two domains, it does not make the grade-level distinctions presented by the framework.

Perhaps one way in which the assessment could be modified to achieve better congruence with the framework would be to weave the idea of science themes throughout more of the assessment. As stated in the framework, "the 1990 Science Framework differs from previous frameworks...in its emphasis on the major themes of science." Although the Science Laboratory Assessment asks the teacher to specify on the teacher questionnaire the theme(s) of the laboratory activity to be observed, the observer is not asked to find evidence that the teacher has presented the theme(s) to the students, either orally, in written materials, or in any part of the lesson. The idea of themes could easily be written into several elements of the **Pedagogy** domain (e.g., *Planning, Prelab, Explanation/Presentation*), and the wording in the description of the *Integrated* element of the **Content** domain could be changed so that the teacher doesn't just know how the activity is related to a major theme of science, but also presents this information to the students.

Two other possible changes to be considered would be to include somewhere in the assessment (e.g., in an element) some reference that (1) the teacher discusses or presents, whenever possible and appropriate, any values and ethics associated with the science activity, and (2) the teacher uses written instructional materials that meet the standards in the framework.

Extent of Coverage of California Standards for Beginning Teachers

Because the Science Laboratory Assessment was developed for the State of California, the developers designed the assessment to cover at least some of the California Standards for Beginning Teachers. FWL staff examined the four components of the assessment--the questionnaire, the domains and their corresponding elements, and the pre- and post-conference questions--to see how well they assess the California Beginning Teacher Standards which define levels of pedagogical competence and performance that California teacher credential candidates are expected to attain (i.e., Standards 22 to 32). As was done in the March, 1990 report, the standards are reprinted below (in italics), along with an analysis of how the assessment components correspond to each standard.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* Both the **Climate** and **Management** domains address this standard. Elements in the **Climate** domain require the observer to look for evidence that the teacher communicates and interacts respectfully with all students, communicates high expectations for student learning and behavior, and that students treat each other respectfully and politely. An element in the **Management** domain asks the observer to find evidence that the teacher encourages and reinforces appropriate student behavior.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* The assessment requires the teacher to plan a 30-45 minute (depending on grade level) science laboratory lesson for observation and to specify on a pre-observation questionnaire the objectives, laboratory activities, student groups, materials and equipment, and safety issues. The questionnaire does not, however, ask the teacher to specify any assessment plans.

Two domains also address this standard. The *Planning* element in the **Pedagogy** domain defines, at a general level, what the teacher's objective(s) should look like (e.g., involve the development or utilization of one or more of the scientific thinking processes), and the *Sequencing* element in the same domain asks the observer to look for evidence that the teacher organizes the laboratory activity in a logical or purposeful manner that allows students to achieve the lesson objective(s). The **Content** domain has an element which asks the observer to find evidence that the teacher uses methods that are related to the objective(s) of the laboratory activity.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities, and materials that are appropriate for students with diverse needs, interests and learning styles.* Three domains and two conference questions address this standard. The **Knowledge of Students** domain asks the observer to look for evidence that the teacher uses instructional strategies and/or activities that are appropriate and challenging for diverse students (e.g., different ethnic, cultural, language, and socioeconomic backgrounds, and disabled students) and students with different student characteristics (e.g., different interests, cognitive and developmental levels, prior knowledge).

In the **Pedagogy** domain, the observer is expected to look for evidence that the teacher knows the students' prior learnings, gives directions, explanations, and presentations at an appropriate level of complexity and difficulty for the students, and provides objective feedback to and asks questions of students regardless of ability, ethnicity, or other characteristics. The *Availability* element in the **Materials/Equipment** domain asks for evidence that the teacher has made provisions for materials to be available to physically disabled students.

In the Pre-Observation Conference, the teacher is asked if s/he designed or modified the activity to make it appropriate for the students' background and interests, and also to describe how the lab activity is related to prior instruction (e.g., which might be construed as prior learnings).

Standard 25: Student Motivation, Involvement, and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* Several domains address this standard: The **Pedagogy** domain asks the observer to find evidence that the teacher provides motivating feedback to all students, and that the teacher's questioning involves as many students as possible; the **Materials/Equipment** domain looks for evidence that, regarding materials/equipment, the teacher has provided easy access and enough so that all students can complete the activity; elements in the **Management** domain ask for evidence that the teacher has structured the laboratory activity so that most of the students are engaged in a laboratory task most of the time, and that the teacher encourages and reinforces appropriate student behavior; and the **Climate** domain seeks evidence that the teacher provides all students with an opportunity to participate and learn.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* The **Communication** domain asks the observer to find evidence that the teacher's oral and written communications are clear and "not vague, ambiguous, or incomplete." Although there is no specific mention of the teacher adjusting the complexity of his/her language to the linguistic abilities of the students, the **Pedagogy** domain asks for evidence that the teacher gives directions, explanations and presentations that are at an appropriate level of complexity and difficulty for the students.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* Two domains address this standard. The **Pedagogy** domain has elements which ask the observer to find evidence that the teacher knows what prerequisite skills and knowledge the students have for an activity (but it does not ascertain how s/he knows), monitors student understanding and work during the activity, and adjusts the lesson or activity as needed. The **Climate** domain asks for evidence that "the teacher communicates high expectations for student learning."

Some of the pre-observation conference questions also ask the teacher what s/he knows about the students' prior knowledge, but again do not ask how s/he assessed this knowledge. In the post-observation conference, the teacher is asked if the objectives were attained by the students, and what is the teacher's plan to assess the retention of these objectives.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* The **Pedagogy** and the **Climate** domains address this standard. For the **Pedagogy** domain, one element asks the observer to look for evidence that the objectives for the activity involve the development or utilization of one or more of the scientific thinking processes (i.e, observing, , communicating, comparing, ordering, categorizing, relating, inferring, and applying). Another element asks for evidence that the teacher asks questions that promote higher-order thinking processes (such as those listed above). The **Climate** domain asks for evidence that the teacher fosters an environment that promotes questioning, problem solving, discussion of error, and evaluation of competing ideas.

A pre-observation conference question also addresses this standard by asking the teacher, "What advanced thinking skills (e.g., comparing, estimating, inferring) will students be encouraged to use or required to apply in order to productively participate in this activity?"

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* The **Climate** domain's four elements address this standard by asking the observer to find evidence that the teacher interacts with all students positively, encourages sharing among students, attempts to instill in students positive attitudes about learning and science, and fosters an environment in which the processes of science are important. One of the **Pedagogy** domain's elements also addresses this standard by asking for evidence that the teacher gives feedback to students that "provides positive rewards, useful information, further motivation, or encouragement to students."

Although none of the domains specifically addresses the promotion of students as independent learners, the focus of the assessment is a science laboratory activity which, by its nature, usually involves some form of independent learning by students.

Standard 30: Capacity to Teach Cross-Culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socio-economic differences.* The **Knowledge of Students** domain asks the observer to seek evidence that the teacher tailors instructional activities for a diverse classroom of students with different ethnic, cultural, language, and socioeconomic backgrounds. As is probably the case with any observation system, however, a teacher's capacity to teach cross-culturally can probably only be demonstrated if the teacher is teaching in an ethnically diverse classroom.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* This standard focuses on a teacher's ability to teach classes which span the range covered by the credential (i.e., grades K-8 or 7-12) or students at two or more ability levels (such as remedial and college preparatory classes). None of the domains are designed to assess this ability. This standard also addresses a teacher's ability to fulfill typical responsibilities of teachers such as meeting school deadlines and keeping student records, none of which are assessed by any of the domains.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* None of the domains assess whether a teacher fulfills his/her obligations as a member of a profession and a school community (e.g., adheres to high standards of professional conduct and engages in collegial relationships).

The extent of coverage by the Science Laboratory Assessment of the California Beginning Teacher Standards is summarized in Table 4.3. The table lists the assessment's domains and questions which address each standard, and also describes the extent of coverage provided.

Job-relatedness

All 29 of the teachers who participated in the RMC assessment pilot test stated that the seven domains chosen for this assessment are relevant to their job of teaching. The pilot test's eleven observers also evaluated the content of the assessment as being relevant to the job of a new teacher of science laboratory lessons, although one observer qualified his answer: "It is only relevant if the teacher gets to see the report." Two observers praised the instrument's relevance as follows:

> *I think it is vitally important to have a method to assess science knowledge/attitude inquiry questioning techniques/lab safety for new teachers.*
>
> *The assessment is relevant to teaching lab science for any teacher, beginning or experienced.*

As was discussed in the March, 1990 report, the job-relatedness of observation systems is strong because such systems almost always entail observing teachers actually teaching in their own classrooms (or to their own students). Moreover, job relevance is a particularly important factor in evaluating different approaches to teacher competence assessment, because professional practitioners and courts of law consider this factor first when they judge the fairness of an evaluation system. As an observation system, the Laboratory Science Assessment offers direct evidence of actual teaching competence. With such an assessment, it is not necessary to make inferences about how well a teacher conducts instruction.

Appropriateness for Beginning Teachers

In this section, the appropriateness of the Science Laboratory Assessment for beginning teachers is discussed from two perspectives: 1) the perceptions of the participating teachers and assessors, and 2) the teachers' performance on the assessment.

TABLE 4.3

EXTENT OF COVERAGE BY THE SCIENCE LABORATORY ASSESSMENT OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Domains and Conference Questions Addressing Standard	Extent of Coverage
22: Student Rapport and Classroom Environment	-Management -Climate	Full
23: Curricular and Instructional Planning Skills	-Pedagogy -Content	Partial
24: Diverse and Appropriate Teaching	-Knowledge of Students -Pedagogy -Materials/Equipment -(Pre-Obs. Conf. #3, #6)	Full
25: Student Motivation, Involvement and Conduct	-Pedagogy, Climate -Materials/Equipment	Full
26: Presentation Skills	-Communication, Pedagogy	Full
27: Student Diagnosis, Achievement and Evaluation	-Pedagogy -Climate -(Pre-Obs. Conf. #5, #6) -(Post-Obs. Conf. #2, #3)	Partial
28: Cognitive Outcomes of Teaching	-Pedagogy -Climate -(Pre-Obs. Conf. #10)	Full
29: Affective Outcomes of Teaching	-Pedagogy, Climate	Full
30: Capacity to Teach Crossculturally	-Knowledge of Students	Partial
31: Readiness for Diverse Responsibilities	-None	None
32: Professional Obligations	-None	None

Perceptions. When asked if they felt they have had an opportunity to acquire the knowledge and abilities measured by the Science Laboratory Assessment, approximately three-fourths (21 of 29) of the teachers responded positively; seven said "no," and one did not respond. Of the seven teachers who replied negatively, four specifically commented that one year is not enough time to achieve mastery of skills and knowledge. This sentiment was also echoed by two of the teachers with positive responses who stated that they were in their second year of teaching and their answers might be different if they were in their first year.

The eleven assessors were also asked if they thought new teachers have had an opportunity to acquire the knowledge and abilities measured by the assessment. Five assessors responded positively, albeit two with qualifications (e.g., "if criteria [are] not too objectively applied"). One of the five stated that a teacher's academic and professional preparation "covers all areas." Another commented, "no problem," because the assessment focuses on minimal proficiency.

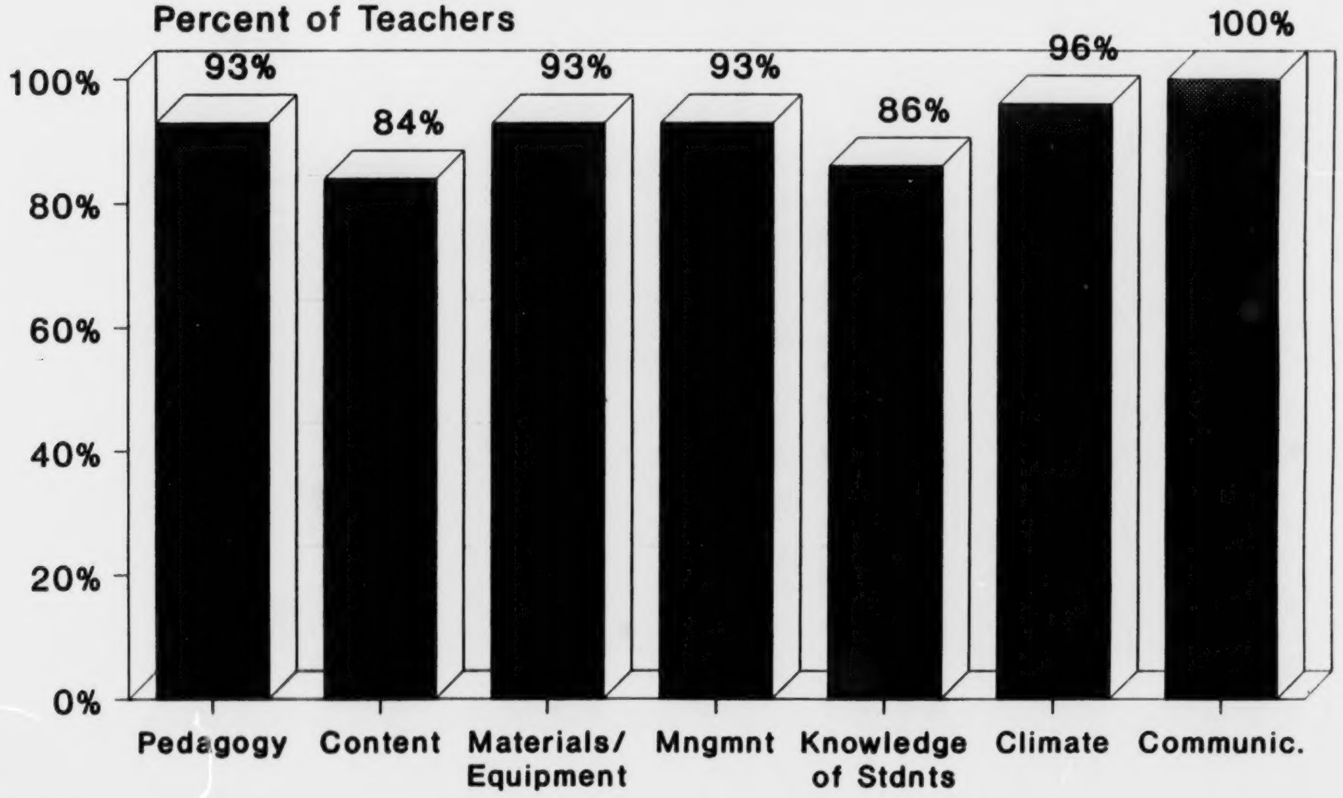
Of the remaining six assessors, one did not respond, two gave answers which were ambiguous, and three expressed a belief that the instrument may be too difficult for a "brand new" teacher (i.e., less than one year's experience).

Performance on assessment. Overall, FWL staff's analysis of the rating results support the majority contention that the new teachers have had an opportunity to acquire the skills and knowledge measured by the assessment. Of the 29 teachers observed, all but two received an overall judgment of passing (i.e., a "2" rating). One teacher received an "X" rating, indicating insufficient information to warrant a judgment, and the other teacher was not given an overall rating. Furthermore, no teacher failed (i.e., received a "1" rating) more than two domains (see Figure 4.3), and at least 18 teachers passed all domains.

Of the seven teachers who failed a domain, however, five were in their first year of teaching. Since 17 of the 29 teachers were first-year teachers, almost one third of the first-year teachers had difficulty in at least one area of the assessment. For three of those five teachers, that area was the **Content** domain.

In fact, of the seven teachers who failed a domain, four failed the **Content** domain. The reasons given for their "failure" tended to fall into two categories: (1) insufficient or missing content, and (2) inaccurate content. For example, a middle school, life science teacher whose laboratory activity was a frog dissection, was given a "1" rating

FIGURE 4.3
Percent of Teachers Receiving a "2" Rating on Each Domain



because the content she presented was "not extensive," but rather was "mostly label the diagram." In addition, the observer faulted her for "little or no discussion of the function of organs in humans vs. frogs." A high school, chemistry teacher who conducted a lab titled "A Reaction with Copper," was given a "1" rating because she did not know the symbol for copper. Her assessor also remarked, "Should not mass anything while hot as it will give results too small." A different error of commission was made by a high school, biology teacher who designed his own lab for an endocrine unit. He was failed because "the relationship the teacher was suggesting (iodine-thyroxine-respiration rate) is not valid." He was also cited for insufficient or missing content: "The relationship between the temperature and respiration rate is valid, but was not actually addressed by the teacher to the students."

As evidenced above, the reasons given for failure, while falling into two categories, were vastly different. One teacher is failed for not knowing the symbol for copper; another is failed for misstating the relationship between an element, an amino acid, and a biological function. Moreover, it is interesting to note that although the **Content** domain was written so that a teacher could be failed for inaccurate content, absolutely nothing was written to suggest that a teacher could be failed for insufficient or missing content. (Revisions were made, however, after the pilot test, to incorporate the concept of sufficiency into the scoring process.)

One last note about content. The fourth teacher who received a "1" rating for the **Content** domain was one of five teachers who were double scored. One of the two assessors who observed this teacher passed her on the domain, and the other failed her. The assessor who failed her described her content as follows:

> *Not accurate because printed materials erroneously confused mass and force leading teacher to the same error. Teacher let students have weights touching bottom of container while reading mass (weight?) on spring scale, and did not recognize this was not an accurate reading.*

The assessor who passed her did not note any of the above problems and described her content as accurate.

Another problem area for teachers--and also for assessors--was the **Knowledge of Students** domain. Of the five teachers who received an "X" rating (i.e., a borderline performance or insufficient information to make a rating), three received the rating in the **Knowledge of Students** domain. These three "X" ratings suggest two possibilities: One, it may be difficult for new teachers with little classroom experience to exhibit the kind of behavior required by the **Knowledge of Students** domain. One of the domain's elements, for example, specifies that, "the teacher tailors instructional activities for a diverse classroom of students with different ethnic, cultural, language, and socioeconomic backgrounds and, when present, disabled students...and each student is challenged at an appropriate level. It is possible that this is a lot to ask of a beginning teacher. Or, as one assessor stated,

> *Some of the domains/elements seem to me to be very advanced teaching skills that most beginning teachers will not yet have acquired (i.e., ability to adjust an individual activity in a variety of ways to meet different student needs).*

A second possibility is that these ratings indicate that it is difficult for an observer to assess this domain unless the observer has a strong knowledge of the students in the classroom. This possibility, however, will be discussed in the next section, "Appropriateness across Contexts."

Although, as indicated earlier, the majority of teachers felt they had an opportunity to acquire the skills and knowledge measured by this assessment, many of the new teachers agreed with the above assessor who thought some domains/elements are harder than others. When asked to name the domains/elements which they thought a new teacher could pass only after two years of experience in the classroom, 15 and 13 teachers respectively named the **Management** and **Knowledge of Students** domains (see Figure 4.4). It is also interesting to note, however, that when asked which of the domains/elements could be passed immediately after student teaching, more than half of the teachers (19 of 29) named the **Content** domain, but all other domains received less than half of the teachers' votes (see Figure 4.5). Looking again at the rating results, the data seems to suggest that 1) the teachers' perceptions of their mastery of science content may be inflated, and 2) their perceptions of the difficulty of passing the **Knowledge of Students** domain may be more on target. Based on the teachers' perceptions of domain/element difficulty, the high rate of passing may also be attributable to the fact that the observations were conducted in the spring when all of the teachers had at least close to one year's experience in the classroom.

Figure 4.4
Domains Teachers Believe Could Only Be Passed W/ 2 Yrs. Classroom Experience

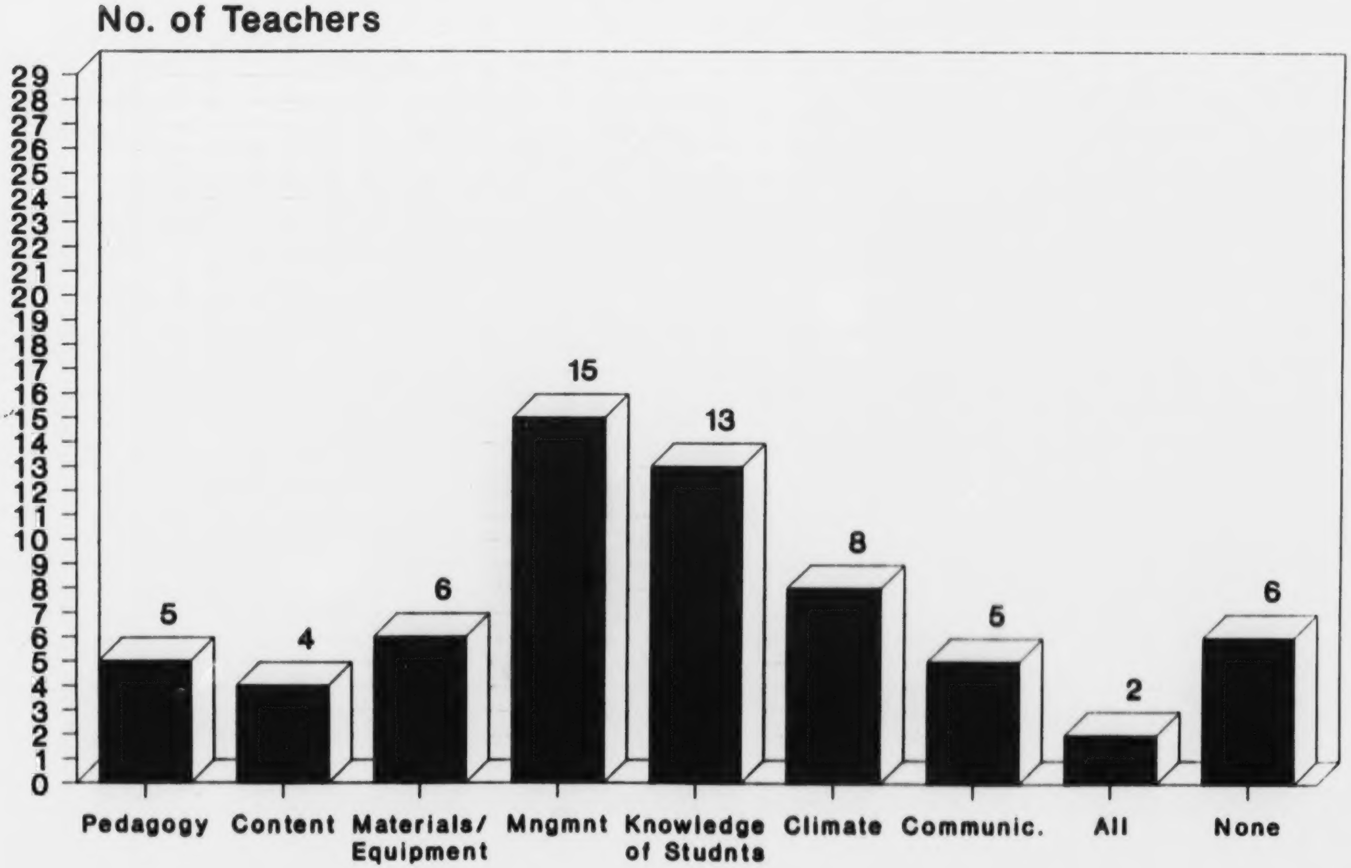
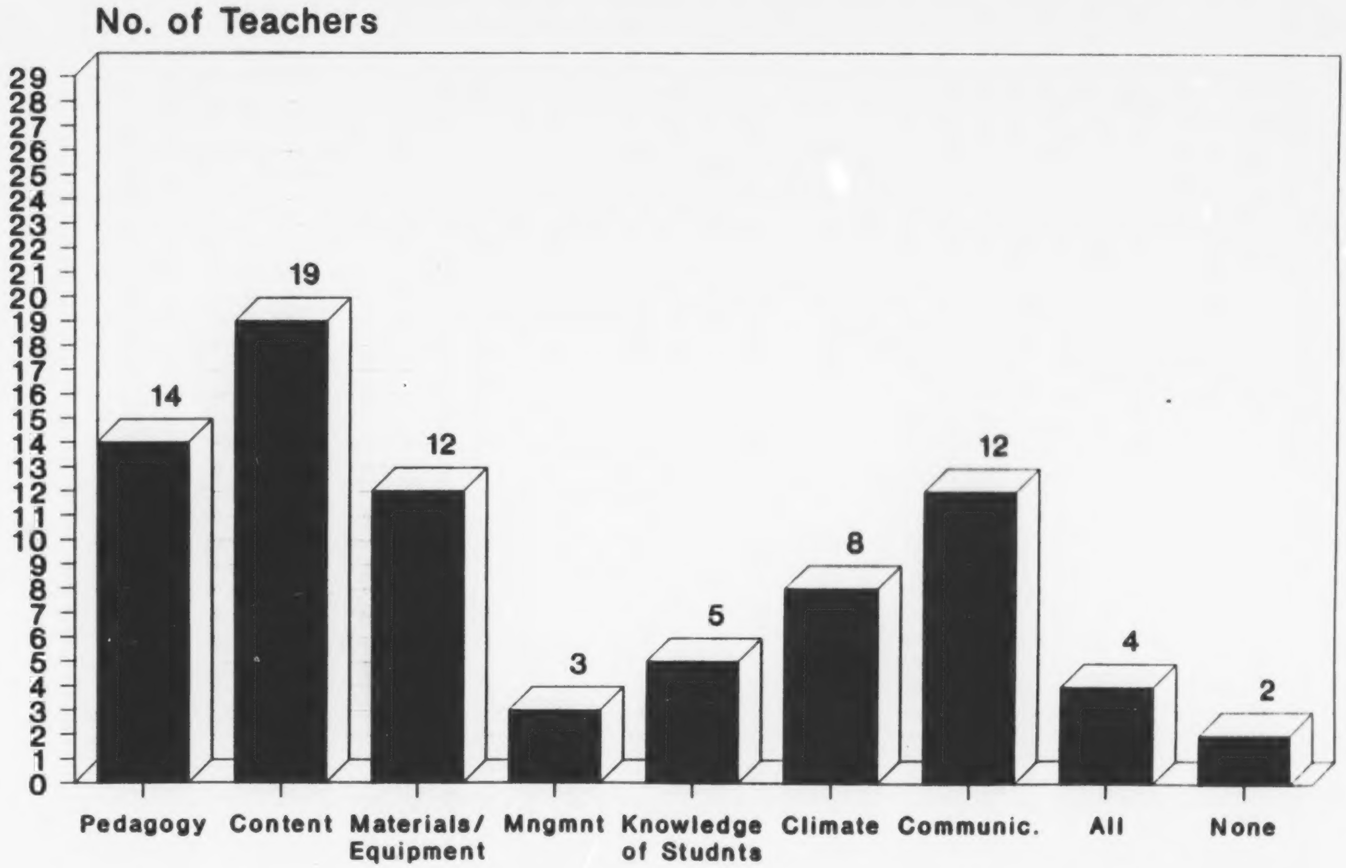


Figure 4.5
Domains Teachers Believe Could be Passed Immediately After Student Teaching



Appropriateness across Contexts

In order to determine if the teachers and assessors felt the Science Laboratory Assessment is appropriate for teachers across contexts, we specifically asked them to comment on the assessment's appropriateness across grade levels, for teachers of diverse student groups, and/or in different school/community settings.

Across grade levels. Approximately 83% of the teachers (24 of 29) felt the assessment is appropriate for teachers across grade levels. Four teachers disagreed and one did not respond to the question. Of the four who disagreed, all were middle or high school teachers who thought the assessment was less appropriate for elementary teachers. Their reasons for disagreement, however, were not compelling. One middle school teacher, for example, commented that "elementary teachers and students would feel very uncomfortable with someone looking over their shoulder." Another middle school teacher stated, "It is more important for a primary school teacher to have good knowledge of students and a positive climate, than to worry about the planning and sequence."

The scoring results do not suggest that elementary teachers are penalized by this assessment. As depicted on Figure 4.6, the elementary teachers performed as well or better on each of the assessment's domains. For every domain but one, all five elementary teachers received a "2" rating. For the **Content** domain, one teacher received an "X" rating.

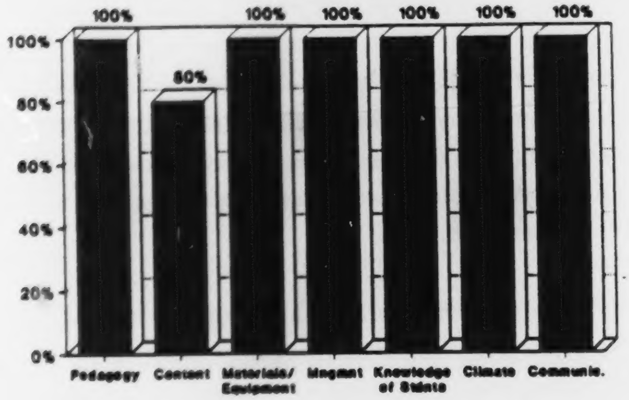
The assessors' comments regarding the elementary teachers' performances also support the idea that the assessment is as appropriate for elementary teachers as for middle and high school teachers. For example, one assessor observed a first-year male elementary science teacher teaching a second-grade class a lab activity involving a comparison of seeds. The following is a comment made by the assessor to explain why she gave a "2" rating to the teacher's performance in the **Pedagogy** domain:

> *The lesson involved exploration and imagination. Students were to use scientific thinking to come up with their own ideas about how seeds might be designed for dispersal. The directions were weak, but the teacher did excellent job of eliciting ideas from students through questioning.*

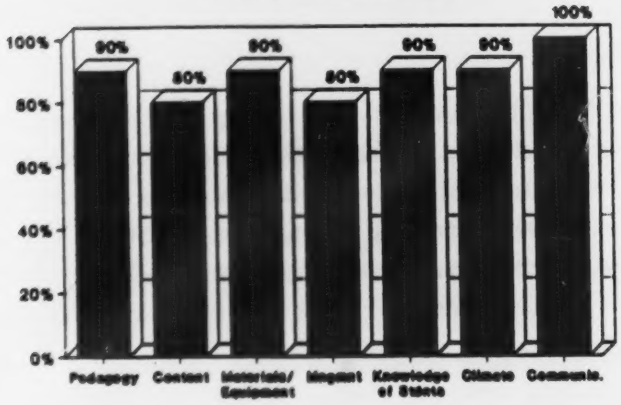
FIGURE 4.6

PERCENT OF TEACHERS BY GRADE LEVEL RECEIVING A "TWO" RATING ON EACH DOMAIN

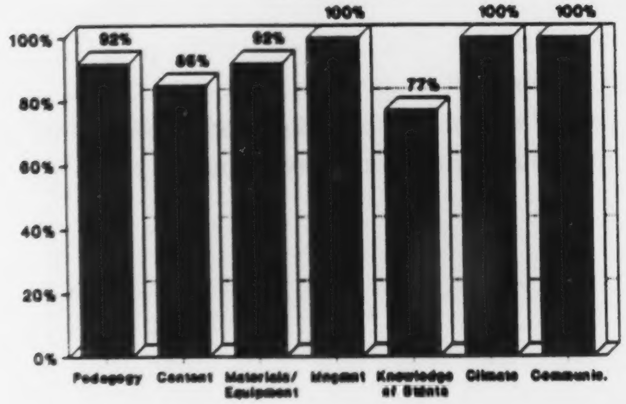
Elementary N=6



Middle School N=10 (1 Missing)



High School N=13



Another elementary teacher, a kindergarten teacher with a Multiple Subjects Credential, was observed teaching an A.I.M.S. science activity called "Huff and Puff," which is part of a larger unit on aeronautics. According to the assessor who observed the lesson, the activity demonstrates that "air is energy and can be used to do work." Working with approximately eight students at a time (an aide worked with the remainder of students on something else), the teacher involved the students in blowing a variety of objects to see how many "blows" it took to move each object three feet. The activity involved problem-solving (i.e., students had to guess how many "blows" each object would require), collecting data, and recording data on a graph. For double-rating purposes, two assessors observed this teacher instructing the activity. Each assessor gave the teacher an overall rating of "2," and each had high praise for the teacher:

> *I observed, to my astonishment, a kindergarten teacher doing exactly what I try to do at the senior high level, using the same skills, the same inquiry methods, and doing it very, very well.*
>
> *Excellent young teacher. Not really her first year. The best inquiry/critical thinking skills type questions I've heard in a long time!*

It should be noted that this teacher was teaching her first year in California, but had taught for two years in another state. It should also be noted that all five of the elementary teachers who participated in the pilot test were either hired as elementary science teachers or had received substantial science training through their district. Thus, when FWL staff agree that this assessment seems appropriate for teachers of all grade levels, we mean to say that it seems to be a fair assessment for those teachers, regardless of grade level, who have been trained to teach science.

Diverse students. The developers of the Science Laboratory Assessment are well aware of the increasing diversity in California's classrooms. As a result, they included in the content of their assessment a domain specifically targeted to assessing a new teacher's ability to work with diverse students. This domain, **Knowledge of Students**, was designed to assess a teacher's ability, within a laboratory setting, to teach (1) students with different ethnic, cultural, language, and socioeconomic backgrounds, and disabled students, and (2) students with different interests, cognitive and developmental levels, and prior knowledge.

Awareness of student diversity is not limited to one domain, however. Observers are also asked to consider a teacher's response to student diversity in other domains. For example, the elements, *Feedback* and *Questioning*, of the **Pedagogy** domain include specific references to student diversity; the *Grouping* element of the **Management** domain asks the observer to find evidence that the teacher has considered the "variable work rate of different students"; and the *Availability* element of the **Materials/Equipment** domain requires that the teacher makes materials and equipment accessible to physically disabled students, when present.

On the surface, then, it would seem that the Science Laboratory Assessment is able to assess a teacher's ability to work with diverse students. But can it? Although approximately 90% (26 of 29) of the teachers felt the assessment is appropriate for teachers of diverse student groups, the assessors were not so quick to agree. Almost half of the assessors named the **Knowledge of Students** domain as the hardest domain to rate, and more than half of the assessors (6 of 11) had serious reservations about using the assessment to assess a new teacher's ability to work with diverse student groups. Said one assessor:

> *I feel the assessment was weakest in this area. Teachers did not seem to have confidence and complex enough skills to really discuss this area and it wasn't always possible to observe needed skills in one observation.*

This assessor also posed the question:

> *Because these teachers are just beginners, is it realistic to expect them to be able to adjust their activities/techniques to meet individual needs and be able to discuss how/why they do what they do?*

Another assessor echoed this sentiment, stating that "the skill/ability to work with heterogeneous groups is the most difficult to learn....it takes time to get good at working with diverse groups."

The majority of assessors, however, did not see the problem as one residing with the teachers, but rather with the assessors themselves. These assessors commented on the difficulty of always being able to recognize the different kinds of diversity among students. For example,

> *Knowledge of Students was difficult to assess by mere observation without questioning the teacher. It is difficult to "see" which students are slower or faster so you can judge if the teacher deals with them differently.*

and,

> *There are places where evidence or notes can be made, yet I found this type of evidence hard to gather, partly because I did not know what students were GATE, LEP, etc. unless I asked the teacher to point them out.*

Similarly, an assessor who pointed out that "behavior problems are easier to identify then LEP or "science shy" types, commented,

> *It's hard to know when the teacher has tailored lessons for students, and then to actually observe that, if you don't know* *who* *the students are.*

In other words, unless an assessor observes a teacher making a major blunder, an assessor's ability to assess a teacher's ability to work with diverse students depends largely on the assessor's knowledge of the students. Without such knowledge, it is, as one assessor stated, "hard to make an educated and informed opinion."

Thus, while the Science Laboratory Assessment was designed to take into account the diversity of California's classrooms, it is questionable whether it is an effective way of assessing a beginning teacher's ability to work with diverse students. Although further study would need to be done to answer that question, FWL staff believes the Science Laboratory Assessment can be commended for putting a focus on a teacher's ability to work with diverse students, and recognizes that such a focus has the potential of improving teachers' skills in this area.

Fairness across Groups of Teachers

A majority of the teachers and the assessors responded positively to the question of fairness of the assessment across groups of teachers (e.g., different ethnic groups, different language groups). Only one teacher gave a negative answer, and this was a Caucasian female who stated that she did not feel qualified to speak for other ethnic/language groups. Of the ten assessors who responded to the question, nine seemed to agree with the assessor who stated:

> *The domains of learning are the same regardless of the characteristics of the teacher.*

One assessor, however, was not sure of the assessment's fairness because, as she explained,

> *Issues of management, climate, and communication will certainly be open to question. It is always possible that an observer may misinterpret or miss pertinent evidence.*

Our very limited sample showed no trends where certain groups of teachers did less well than other groups. For example, of the five teachers who are non-Caucasian, non-Hispanic, only one did not receive a "2" rating for all seven domains (one teacher received a "2" rating for six domains). Of the seven teachers who received one or more "1" ratings, we observed no clear patterns regarding age, gender, etc.

Appropriateness as a Method of Assessment

In addition to evaluating the appropriateness of the Science Laboratory Assessment for beginning teachers, and its appropriateness across contexts and groups of teachers, the teachers and assessors were asked to evaluate the appropriateness of the method of assessment, and to compare it with other methods of assessment which they have experienced.

Appropriateness. The teachers were asked if they thought this type of assessment (i.e., classroom observation of a science type of laboratory activity) is an appropriate way of assessing 1) general teaching skills, and 2) skills in teaching laboratory science. Their answers to both were positive, with 76% (22 of 29) and 79% (23 of 29) of the teachers replying "yes" respectively. The assessors were even stronger in their affirmation, with 90% (10 of 11) and 82% (9 of 11) saying "yes" respectively to the assessment's appropriateness in assessing general teaching skills and skills in laboratory science. One assessor remarked,

> *This is an organized way to take a look at science teachers and see what they actually do in a way that helps delineate excellent practices as well as practices that are missing, yet needed.*

Another assessor, however, disagreed that the assessment is an appropriate way to assess the science teaching skills of beginning teachers. As she explained,

> *I think a lot of first year teachers shy away from a lot of lab, hands-on activities because they are 1) unfamiliar with what is available in the school/community, 2) labs take time to set up, and 3) management is important in a lab and not well developed your first year of teaching.*

Comparison. All of the teachers were asked to compare the Science Laboratory Assessment with other assessments with which they have been evaluated (e.g., multiple-choice exams such as CBEST and NTE Specialty Area Tests, classroom observations during student teaching) in terms of its ability to assess teaching competency. Approximately 70% of the teachers (21 of 29) commented favorably about the assessment, many stating that the assessment is better than the NTE and/or CBEST tests. Almost 20% of the teachers (5 of 29) said the assessment compared favorably with the classroom observations they had during student teaching. Some teachers commented as follows:

> *Much superior to test of subject areas knowledge. Compares favorably with student teaching observations.*

While there were no distinctly negative comparisons, some teachers did describe weaknesses of the assessment. For example, four teachers said they did not like the assessment or could not compare it with others because they did not receive any feedback; two teachers stated that this type of assessment makes a first-year teacher very nervous because "you are put on the spot to perform"; and one teacher, who commended the assessment for "evaluating a lesson (which CBEST and NTE do not)" expressed concern that, in this assessment, the observer "becomes a stenographer and spends an excessive time writing down as much as possible to find each domain and element" so that there is the possibility that "the big picture is lost while the details are haplessly pursued."

Finally, four teachers said the assessment would be best if used in conjunction with other methods of assessment.

Assessment Format

Although the classroom observation is a traditional method of teacher assessment, the Science Laboratory Assessment breaks new ground because it was designed to focus on a teacher's performance during a particular activity (i.e., a laboratory activity) in a particular subject area (i.e., science). Thus, whereas all classroom observation systems are relatively easy to administer because they require minimal materials (e.g., paper and pen for the assessor), the Science Laboratory Assessment may be more difficult to administer because it requires the observer to assess a teacher's performance while s/he is engaged in a specific activity that, by definition, is more student-centered than teacher-centered. Moreover, the Science Laboratory Assessment is not a checklist, but requires the observer to collect evidence during the observation by constantly writing down exactly what the observer sees during the lesson. In addition, the assessment requires the observer to categorize the evidence at the same time s/he is collecting it (i.e., writing it down). The assessment's analysis process also entails much more writing than traditional systems, and, perhaps, more careful codification.

Still other format issues to consider are 1) the Science Laboratory Assessment, like all observation systems, cannot easily be administered to groups of teachers because its format requires one assessor observing one teacher at a time, and 2) the assessor must be able to travel whatever distance is necessary to observe at the teacher's school site (or science laboratory setting) because that is where the assessment takes place. As was mentioned in the March 1990 report, these format issues pose a formidable challenge in the state of California.

In this section, the format issues which can be more easily addressed will be discussed. These include the clarity of the assessment's preparation materials for the teachers, the clarity of the pre- and post-observation conference questions, and the clarity of the documentation and rating forms and process used by the assessors.

Clarity of the Teachers' Preparation Materials

To prepare for the assessment, each participating teacher was asked to 1) read an orientation handbook which described the content (i.e., domains, elements, and indicators) and format of the assessment, and 2) complete the Pre-Observation Questionnaire which was found in the handbook. Although only 55% of the teachers (16 of 29) said they read the orientation handbook carefully, 93% (27 of 29) said that the handbook clearly described the aspects of teaching being measured by the assessment. Some teachers praised the handbook as "very good" and "very clear and understandable."

When asked to offer suggestions for improvement, several teachers commented on the size of the handbook, describing it as "massive" or "too long," and suggested it be simplified. Only one teacher, however, offered a suggestion as to how it might be simplified, citing the existence of "several redundant sheets in the examples section." Other suggestions for improvement included 1) more warning as to the amount of paperwork (i.e., the Pre-Observation Questionnaire) they had to complete before the observation, and 2) more specific labeling of the "seven pages of terminology" (i.e., the domains, elements, and indicators) as the assessment content.

Reviewing the orientation handbook sent to teachers, FWL staff did not find the materials to be massive or too long, but did feel that they could be better organized. A Table of Contents could be provided so as to alert the teacher to the specific contents of the handbook, and the seven pages describing the assessment content could be immediately proceeded by a brief introduction which clearly states that the information which follows (i.e., the domains, elements, and indicators) are those things which the observer will be looking for during the observation. In the examples section, there are three examples, each of which includes two pages (pgs. 7 and 8) that are exactly the same, and, therefore, could be considered redundant. FWL staff does not believe, however, that the elimination of these pages would improve the materials, because the pages are reference materials to be used in conjunction with some of the questions on the preceding page (i.e., p.6).

The six-page Pre-Observation Questionnaire that the teachers had to complete consisted of a variety of questions about the class to be observed, as well about the laboratory activity to be taught. In addition to the questions, the teacher is asked to fill out a chart describing the objectives, activities, student groups, materials and equipment and safety issues for the lab activity. The questionnaire is then read by the assessor before the observation and the pre-observation conference.

When the teachers were asked if they had any difficulties completing the questionnaire, 93% (27 of 29) said "no." Of the two teachers who said "yes," one claimed she did not receive the questionnaire, and the other stated she had difficulty because she was "unaware of the list of science themes" and that "this was information that I should have been aware of through my teacher education classes." (The list of science themes (e.g., energy, environment, stability, evolution) is included as part of the questionnaire to help the teacher answer question #14 which asks, "Which scientific theme(s) best pertains to your laboratory activity?")

Although the majority of teachers said they did not experience any difficulties with the questionnaire, it should be noted that some of the assessors thought otherwise. An assessor who observed three teachers, two of whom had not completed the questionnaire when he arrived, commented that the teachers "have problems stating objectives if the lab is part of regular curriculum material." Another assessor remarked that at least three of the four teachers she observed seemed "overwhelmed" by the materials they were given, and that the questionnaire "seemed wordy, too lengthy and burdensome for these people to deal with."

One other piece of evidence that suggests that the teachers may have had problems with the questionnaire even if they did not say so is the amount of time it took the teachers to complete the questionnaire. Although the majority of teachers who specified their time said they needed from between 20 and 40 minutes to complete the questionnaire, one-fourth of the teachers (7) said they needed more than 40 minutes, and three of those teachers needed one hour or more. The longer time periods taken to complete the questionnaire may be another indication that some of the teachers had difficulty or were at least unfamiliar with some of the tasks on the questionnaire.

Although FWL staff acknowledges that a six-page questionnaire could be considered lengthy, we feel that all of the information on the questionnaire is important for an observer to know before conducting an observation. In addition, even though the task of describing (on a chart) the objectives, activities, student groups, materials and equipment, and safety issues for the laboratory activity may be time-consuming, it is a task which, we believe, all science teachers should be able to complete in a competent manner.

Clarity of the Conference Questions

The Science Laboratory Assessment includes pre- and post-observation conferences conducted by the observer with the teacher. The pre-observation conference, which is conducted after the observer has read the teacher's questionnaire, consists of 15 questions, and the post-observation conference consists of 8 questions. The data gathered from both conferences is used in the analysis and rating process of the assessment.

Both the teachers and the assessors were asked if they had any difficulties with the two conferences, and almost all of them said "no." Only two teachers and three assessors stated they had difficulty with either the Pre- or Post-Observation Conference. One of the two teachers expressed displeasure with not receiving any feedback after the Post-Observation Conference, and the other teacher complained of "too little time" allotted for the Post-Observation Conference. For the three assessors who had difficulty, lack of time was also an issue. One assessor commented:

> *I found the pre-ob [conference] took a minimum of 30 uninterrupted, on-task minutes. Most of my teachers only allowed 20 and were doing other things (e.g., supervising a class or writing up plans) at the same time.*

Although the developers of the assessment recommend 30 minutes for the pre-observation conference, the teachers were usually scheduled for a 20-minute conference. For some of the teachers, 30 "free" minutes were hard to find. For example, if the teacher was scheduled to be observed in the afternoon, the pre-observation conference usually took place during the teacher's lunch hour. Since many lunch hours are closer to 40 minutes than an hour, teachers were not asked to give up the majority of their lunch hour for the conference. Almost half (5) of the eleven assessors, however, indicated that, on average, they needed 30 to 40 minutes to conduct the Pre-Observation Conference, and more than half (8) needed, on average, at least 20 minutes to conduct the Post-Observation Conference.

When asked if there were any conference questions with which the teachers consistently had difficulty, approximately three-fourths of the assessors (8 of 11) said "no." Three of the assessors disagreed, naming pre-observation conference questions #4, #6, #8, and #9, and post-observation conference questions #2, #3, and #4 as sources of difficulty (see Appendix B). One assessor explained that #6, which in part asks, "What do students already know about this topic?", was difficult because the teachers "cannot say with accuracy what students already know about the topic." Another assessor commented that #8 and #9, which ask the teacher about future instruction, broad goals, and linkage between concepts, were difficult because the teachers he observed "generally indicated they didn't think about broader goals, the bigger picture, and connections."

The assessors were also asked if they thought any of the conference questions could be eliminated or collapsed. Although four assessors said "no," and one assessor didn't respond, six of the assessors suggested changes to the pre-observation conference questions. Among these six assessors, there was some agreement as to the questions which should be changed, but there was not always agreement as to how they should be changed. For example, the question which received the most suggestions for change was #6 which reads, "What prior instruction have you implemented related to the lab activity? What do students already know about this topic?" The three assessors who targeted this question, however, each suggested a different change: 1) eliminate the question, 2) eliminate the second part of the question, and 3) collapse the question with #7 to form a new question.

Altogether, eight of the 15 pre-observation conference questions were targeted for change by one or more of the assessors. Table 4.4 shows the questions recommended for change, the number of assessors who wanted to change the questions, and the changes suggested. FWL staff concur with the assessors who believe the assessment instrument would be improved if (1) changes are made to the pre-observation questions, and (2) the focus of the changes should be on improving the clarity and reducing the complexity of the questions. To this end, we recommend that some of the questions be collapsed (i.e., #6 and #7; #4 and #10), and that #9 should be into two questions or the second part should be eliminated. We also recommend that if the pre-observation conference can be shortened by eliminating or collapsing questions, it should be. Careful review of all the answers given by the teachers to the pre-observation conference questions could provide useful insights as to how the questions can best be changed.

Clarity of the Forms and Process for Documentation and Rating

In order to document and rate a teacher's performance, the format of the Science Laboratory Assessment requires the assessor use three forms: 1) the Guided Note-taking Form, 2) the Documentation Sorting Record, and 3) the Summary Report Form. Because these forms were not provided to or used by the teachers, the teachers were not asked to evaluate them; hence, there is no teacher feedback included in the discussion of the forms.

TABLE 4.4

SUGGESTED CHANGES TO PRE-OBSERVATION CONFERENCE QUESTIONS: SCIENCE LABORATORY ASSESSMENT

Question	No. of Assessors Wanting Change	Suggested Changes
1. I have reviewed your Questionnaire. Is there anything on it you need to change before we continue?	1	-eliminate
4. Explain the scientific concepts and/or skills you are teaching in this lab activity.	2	-collapse with #10
5. What are some of the incorrect preconceptions that students may have that relate to this activity? (pause) How do you plan to address these during the lesson?	1	-collapse with #9
6. What prior instruction have you implemented related to the lab activity? (pause) What do students already know about this topic?	3	-eliminate; eliminate 2nd part; collapse with #7
7. Have you provided previous instruction to ensure that students have the technical skills (e.g., students know how to use a voltmeter) requisite to the successful completion of this laboratory activity? If yes, was this provided recently? If not, what techniques have you employed to provide you evidence that students are ready to use the required processes and technical skills?	2	-eliminate; collapse with #6
8. What instruction are you planning to do in the future related to this activity?	1	-collapse with #9

TABLE 4.4 (continued)

SUGGESTED CHANGES TO PRE-OBSERVATION CONFERENCE QUESTIONS
SCIENCE LABORTORY ASSESSMENT

Question	No. of Assessors Wanting Change	Suggested Changes
9. What is the relationship or contribution of this laboratory activity to the broad goals for the students' learning? (pause) Does it provide linkage from one concept to the next, or is part of a continuing direction within one major concept? If yes, please explain.	2	-make into two questions; collapse into #5 and #8
10. What advanced thinking skills (e.g., comparing, estimating, inferring) will students be encouraged to use or required to apply in order to productively participate in this activity?	2	-collapse with #4

Guided note-taking form. As described at the beginning of this chapter, the Guided Note-taking Form (from here on after referred to as GNF) is the form used by the assessor to record that which is seen and heard during the observation. The form is divided into seven spaces, each space corresponding to one of the assessment's seven domains. During the observation, the assessor is expected to simultaneously record and categorize the evidence/notes from the observation into the appropriate space, (i.e., the appropriate domain). On average, an observer will record evidence and notes on 12-15 of these forms during a single observation.

During the assessment training, the assessors expressed much frustration using the forms. Although many had experience scripting observations (i.e., writing down everything they saw/heard during an observation in a chronological manner), none had experience scripting and categorizing information at the same time. The trainers acknowledged the frustration, but instructed the assessors to continue using the forms in the hope that with practice the assessors would become more comfortable and proficient in using the forms.

In fact, many of the assessors did become more comfortable and proficient using the form with practice, as indicated by assessor comments such as these:

> *I was confused at times, but felt more comfortable the more I did it.*
>
> *It got easier to use the GNF with experience.*

Nevertheless, when asked if they had any difficulties with choosing the category in which to record evidence, 82% of the assessors (9 of 11) said "yes." For almost all of these assessors, the difficulties were a result of 1) not being clear on what the domains and elements meant, and 2) not being clear as to what to do with evidence that, in their opinion, fit into more than one domain. For example, in the comment below, an assessor explains how he was unclear about the meaning of elements within and across domains:

> *In the Pedagogy domain I had difficulty distinguishing between the elements, "Directions" and "Explanations/Presentations." I also feel that Pedagogy's element, "Questioning," overlaps Climate's "Inquiry" element.*

Similarly, in the next comment, another assessor describes his difficulty choosing in which domain and elements to record evidence:

> *I had problems when dialogue would fit into more than one category such as Climate's "Interactions with Students," and Pedagogy's "Monitoring and Adjusting" or "Feedback"....*
> *Many times I would put the dialogue into both.*

Clearly, as was discussed in the section, "Assessors and Their Training," the assessors would have benefitted from more training in the meaning of the domains and elements, as well as what to do if evidence falls into more than one category on the Guided Note-taking Form. Indeed, because the training did not directly address this issue, nor was it addressed at all in the Assessor's Handbook, it is hard to evaluate whether the difficulties described above in using the Guided Note-taking Form are inherent to the form itself, are a result of the training, or both.

In addition to the difficulties described above, approximately one third (4 of 11) of the assessors described difficulties that had little or no relationship to the training. One assessor, for example, emphatically expressed difficulty not with choosing the categories, but with trying to observe, write, and categorize the evidence simultaneously:

> *It is very hard to move around the class and see the specific categories, and record exact quotes.*

Another assessor also expressed difficulty with the guided note-taking format, commenting that, because there is "no requirement to note what the teacher does not do, if the scripting is not complete, there is no way to know." In other words, because the GNF format requires the observer to categorize evidence as it is observed instead of scripting the entire lesson in complete chronological order, the GNF format misses the flow and continuity of the lesson and thus increases the chance that the observer may miss information about what the teacher has not done at a particular point in the lesson.

Finally, other assessors found fault with the form itself, and offered suggestions, such as the following, as to how the form could be improved: reduce the size of the margin and the amount of information (e.g., title of form, slots for names, time, and date) at the top; and add to the top of the form a slot for the setting (e.g., whole-group, small-group).

Should the Guided Note-taking Form be retained as part of the Science Laboratory Assessment, FWL staff agrees that the form would be improved by following the above assessors' suggestions.

Documentation sorting record. The seven-page Documentation Sorting Record (DSR) is the second step in the Science Laboratory Assessment's documentation and analysis process. After using the Guided Note-taking Forms to categorize the evidence/notes from the observation by domain, the observer uses the DSR to further sort the evidence/notes by element. The observer also uses the DSR to sort by element all of the information collected from the pre- and post-observation conferences and the questionnaire. As there are a total of 33 elements, completing the DSR requires a considerable amount of work by the observer.

As with the Guided Note-taking Forms, the assessors were asked to describe any difficulties they may have had using the Documentation Sorting Record. The major difficulty, cited by all but two of the eleven assessors, was that the DSR process, is "time-consuming" and "very laborious." Assessors claimed the DSR took them from between two and five hours to complete. An assessor who completed four observations remarked,

> *Frankly, this part of the process I found an absolute "bear." It takes a long time, 2-4 hours, to get through the sorting record.*

One assessor pronounced the DSR to be the "weakest link" in the assessment process "because of length and consequence of time."

To reduce the time, some of the assessors suggested changing the DSR process. A couple of assessors advocated omitting the DSR completely and just relying on the GNF. As one assessor explained, "I felt that my original notes on the GNF were legible and clear so that I could go directly to the Summary Report Form." Two other assessors suggested that the DSR be used more selectively, such as to note "only critical (+ or -) evidence" or to use the form only when there is a "potential 1 rating in any domain and a possible overall rating of 1."

While FWL staff agrees that the DSR is a time-consuming process, we do not think that there should be total reliance on the GNF because many assessors' notes are not legible and clear during this step of the process. In fact, it is probably unrealistic, if not unfair, to ask an assessor to not only try to script an observation, but to categorize the evidence/notes as they are recorded, and to do all of this in a clearly, legible manner. Assessors often use a personal shorthand during this process which enables them to capture more evidence/notes. If an assessor was required to always write in a clear and understandable manner at this step of the process, there is a good chance that the assessor, while focusing on legibility, will miss evidence.

For much the same reason, FWL staff believe that the DSR should not be used for only some evidence or certain ratings. An independent reader should be able to read evidence pertinent to all ratings (domains and "overall"), and this may not be possible if the assessor's writing is not legible on the GNF. Furthermore, if an assessor only writes "critical" evidence, there would have to be a clear understanding of the meaning of the word "critical," and this would likely add another subjective component (i.e., the assessor's judgement of what evidence is critical) to the assessment.

Another difficulty, cited by several assessors (3 of 11) and which perhaps contributed to the length of the DSR process, was that of deciding how and where to write the evidence. As suggested by the following assessor, this difficulty may have partly been a result of the training:

> *I had difficulty deciding what to put down and how to get it down on paper. The concept is clear but I didn't feel the training we had was complete enough to enable me to do this step comfortably.*

FWL staff agrees with this assessor that more training in sorting and writing up evidence was needed, and believes that such training would greatly reduce or eliminate the difficulty described above.

FWL staff further believes that consideration should be given to revising the Documentation Sorting Record so that it is more than just a place to recopy evidence/notes and to read all the evidence/notes together. It does not seem worth two hours of an assessor's time to basically recopy notes. The DSR should be revised to somewhat resemble the second step of the Classroom Competency Instrument, an assessment instrument pilot tested last year. That is, the second step in the documentation process would require the assessor to not only sort evidence by element, but to also differentiate whether the evidence was positive or negative. Furthermore, the way in which the evidence is written could be specified so that 1) the assessor does not have to write all evidence but only that which best exemplifies (positive and negative) the elements, and 2) there is some consistency among assessors' write-ups. Such revisions, we believe, would greatly enhance the assessment process as a whole and the DSR in particular.

Summary report form. The Summary Report Form constitutes the third and final step in the Science Laboratory Assessment's analysis and rating process. On this two-page form, the assessor records a rating of the teacher's performance for each domain and assigns an overall rating to the teacher's performance. The assessors are asked to choose between two possible ratings: a "2" rating indicating minimal competency, and a "1" rating indicating a lack of minimal competency. If a choice between the two ratings can not be made because of a borderline performance or a lack of sufficient information, then the assessor gives an "X" rating. After making the rating, the assessor writes three or four summary remarks corresponding to that domain and rating (or the overall rating).

The assessors were asked if they had any difficulty with 1) recording summary remarks for each domain, 2) assigning a rating for each domain, and 3) assigning an overall rating for the teacher. An overwhelming majority said "no" to each of the above. In fact, only one of the eleven assessors expressed any difficulty recording summary remarks, only one had difficulty assigning a rating for each domain, and no assessor had difficulty assigning an overall rating. Of the assessors who did experience difficulty, one was not clear on the difference between a minimally acceptable and not acceptable performance; the other was not clear on how to write the summary remarks. The latter explained:

> *My concern was that I was missing something important that should have been picked up or that I was somehow mishandling the evidence.*

This assessor added that it would have been helpful during training to have had several examples of how different observers write summary remarks.

The assessors were also asked to suggest which domains, if any, should receive more/less weight when considering the overall rating of the teacher. Of the eleven assessors, three felt none of the areas should receive more/less weight, and two assessors did not respond to the question. The answers of the remaining six assessors are listed below, together with the number of teachers who gave them.

Most Emphasis		**Least Emphasis**	
Pedagogy	(4)	Knowledge of Students	(3)
Content	(4)	Climate	(2)
Management	(4)	Communication	(2)
Materials/ Equipment	(3)	Materials/ Equipment	(1)
Climate	(2)	Management	(1)
Communication	(2)	Pedagogy	(0)
Knowledge of Students	(0)	Content	(0)

Some of the assessors explained why they thought certain domains should receive less emphasis. For example, one assessor thought the **Knowledge of Students** and **Climate** domains should receive less emphasis because "they are really included in Pedagogy, Management and Communication." This thought was echoed by another assessor who suggested the **Knowledge of Students** domain receive less emphasis because "beginning teachers have difficulty sorting this out from their overall pedagogy and management." One assessor proposed that the **Management** and **Communication** domains receive less emphasis because "they develop after your first year of teaching." Finally, the assessor who nominated the **Materials/Equipment** domain for less emphasis did so because the "safety aspect of a teacher's handling of materials cannot be determined in some observation activities."

FWL staff's analysis of the assessors' Summary Report Forms raised several concerns, the majority of which revolved around the assessors' summary remarks made for each domain. Focusing on the Summary Report Forms for the five teachers who were double-scored, we found tremendous differences in the summary remarks written by the assessors. For example, for the high school biology teacher who instructed a lab on proteins and oils in the digestive system, the two assessors who observed him each gave him a "2" rating in the **Materials/Equipment** domain. Their summary remarks corresponding to this rating, however, were vastly different. One assessor wrote:

> *A good deal of work done to organize materials for the lab.*

The other assessor's remarks were as follows:

> *The teacher provides verbal safety instructions and points out location of safety and clean-up items. The teacher monitors student use of chemicals during the lab and has prepared and allocated materials so as to save time during the lab. Students might be given more responsibility in labelling tubes for the lab thereby decreasing prep time for teacher. All students assist in an orderly clean-up.*

Although both assessors agree that the teacher is at least minimally competent (i.e., merits a "2" rating) in this domain, we get two different pictures of the teacher's competency when we read the two assessors' summary remarks. The first assessor's remark gives limited information about the teacher's performance in this domain, and equally limited information to support the "2" rating. What does the assessor mean by "a good deal of work"? What about the teacher's performance in the other elements of the domain--e.g., Did the teacher set the materials up safely? Did the teacher and students use the materials safely? Were the materials available to all students? The first assessor's summary remark does not really support or explain why the teacher received a "2" rating. The second assessor's comments, however, give a much richer description of the teacher's performance and competency, and offer support and explanation of why the teacher received a "2" rating. Although the second assessor could be faulted for including a suggestion about how the teacher could reduce prep time, the second assessor's remarks seem preferable to those of the first.

To further illustrate differences in summary remark write-ups, the following are two assessors' summary remarks addressing a Kindergarten teacher's performance in the **Content** domain (both assessors gave the teacher a "2" rating):

First assessor:

> *Teacher does this lab as part of AIMS unit on aeronautics. It shows that air is energy and can be used to do work. Good lab to show that concept at this level. Integration with other units/subjects was weak.*

Second assessor:

> *I noted that the teacher was able to relate easily to previous and future direction "Scientists today," "role playing yesterday," "how far can you," and "lift."*

Of the first assessor's four summary remarks, only one directly addresses any of the three elements (i.e., *Integrated*) of the **Content** domain. The other three remarks are more descriptive of the lab than of the teacher's performance. The second assessor offers only one remark, also addressing only one of the domain's elements (*Integrated*). While this assessor's remark is not especially clear, it does include verbatim examples in support of the assessor's remark.

Thus, while the assessors expressed no difficulty with writing the summary remarks on the Summary Report Form, there is very little consistency among assessors as to what is written and how it is written. Furthermore, more often than not, the summary remarks do not seem to summarize the teacher's performance or adequately support/explain the rating given. As presently constructed, the RMC rating process is essentially a pass/fail system which does not provide information that differentiates among teachers who are at different levels of performance within the domain. Thus, without major revisions, the state could not use this instrument to increase teachers' competencies because there is no basis upon which to do so.

FWL staff suggests that the rating process of this assessment undergo extensive revision so that there is (1) a clearer picture of what a "2" performance looks like, and (2) the instrument could be used to increase teachers' competencies (e.g., through staff development). One possibility for revision is suggested by the state's 1990 Science Framework. Taking the state's criteria for adoption of instructional materials as a basis, the RMC assessment's rating process might be revised to include at least three domains, each of which is weighted (which may be done with points). The first domain, for example, could be **Content** and it would carry the most weight (or points). Within this domain, there would be at least four elements which cover accuracy, depth, use of themes, and process. Each of the four elements could also be weighted (e.g., accuracy would be more important than depth). The other two domains could be **Presentation** and **Pedagogy**, both of which might carry the same weight. Presentation would include elements which address, for example, communication, attitudes toward science, explanation/presentation, and questioning. Pedagogy would include elements which might address grouping, feedback, student engagement, and knowledge of students. The Science Laboratory Assessment's domain of **Materials/Equipment** might serve as a fourth domain, or it could be included in the other domains (e.g., in the **Content** domain under the *accuracy* element--materials and equipment are handled in a correct manner by teacher and students). Many of the assessment's other elements could also be subsumed under the above domains.

With such a rating process as described above, the assessor would be firmly guided in making his/her rating decision and the rating results would more likely show greater differences among teachers' performances. While we can not advocate such a process without further and extensive study, we do recommend that strong consideration be given to reducing the number of domains for which ratings are given, to weighting the domains (especially the **Content** domain), and possibly to weighting the elements. Whatever revisions are made or considered, the end goal should be to produce an assessment in which a "2" rating, for example, is meaningful and consistent across assessors and teachers.

Cost Analysis

Based on our experience pilot testing both this version of the Science Laboratory Assessment and the Connecticut Competency Instrument (CCI) in 1989 we have outlined in this section estimates for administering and scoring this laboratory observation assessment and summarized costs for the development and pilot testing of this prototype. These costs, however, should be taken as only preliminary estimates for what costs would be incurred if an assessment like this were to be further developed and modified for implementation on a wide scale.

Administration and Scoring Cost Estimates

Assessor time and costs. Administering this assessment requires a trained observer-assessor to (a) prepare and arrange for the assessment, (b) review the pre-observation questionnaire form, (c) conduct the pre-objective conference, (d) conduct 30- to 45-minute observation, (e) conduct a post-observation conference, and (f) summarize the evidence/notes taken during the observation and rate the teacher's performance. These activities take approximately 4-6 hours for each observation. Thus, using an hourly rate of $20 per hour implies that it will cost approximately $100 per observation to conduct this assessment.

Training costs for assessors. The training for this assessment consisted of one home-study day and two days of group training. As related earlier, we do not believe this is sufficient time to train assessors to reliably and validly score this assessment. At a minimum, the training should be extended by one day, and we believe that it will ultimately need more like the five days used for the CCI training which also has a two-day follow-up session. For estimates here, we will assume that the training will take four days with no follow-up training needed. If each assessor-observer conducts 30 observations each year for five years, we can distribute the training costs over 150 observations. Reimbursing the assessors for the four days of training at $20 per hour would add about $4 to the cost of each assessment.

Other costs. Other costs include those associated with the telephone, duplication, postage, and travel. Travel could be expensive in California unless regional assessors were used. Estimating costs of these activities or ingredients would depend in large part on the manner in which the system was ultimately designed and how costs were apportioned. Using a figure of $30 per assessment for these activities would assume minimal travel costs, based on our experience from pilot testing.

The above estimates imply that the costs for administering and scoring each assessment could be approximately $134. This figure should be taken as only an initial and rough estimate. Actual costs would depend largely on the rates and methods for paying assessors, whether the assessments were administered with local or centrally based assessors, and the degree to which the training and administration times for the final assessment were within the estimates used here.

Development and Pilot Testing Costs

Although the development and pilot testing of the Science Laboratory Assessment was much closer to a research and development stage than an implementation stage, it still may be helpful to report the costs associated with developing and pilot testing this prototype.

Development of this prototype, as described earlier, drew substantially from the experience of other teacher performance assessment systems such as the CCI. Thus, the development for this assessment benefitted from the prior development and materials available from these earlier efforts. Costs for Development and for Pilot Testing are outlined in Table 4.5 in terms of the developer and pilot test staff time, consultants to the developers and in the pilot testing (e.g. consultants in the pilot testing include costs for reimbursing the teachers and assessors participating in the pilot test), travel, and other direct costs for items such as phone, duplication, facilities, etc.

Cost Summary

The experiences from pilot testing a limited number of the Science Laboratory Assessment provides some initial estimates that might be expected with developing and implementing an assessment of this type. The development and pilot testing costs could be reduced with a larger scale and more advanced stage of development. Similarly, the costs for administering and scoring will depend on the number of teachers to be assessed, the location and costs associated with training and supporting assessors, and the methods used to pay these costs. For example, using retired teachers versus the use of practicing teachers as assessors and providing substitute teachers, would likely result in different costs.

Technical Quality

This section briefly discusses three technical issues related to the Science Laboratory Assessment -- development, reliability and validity.

Development

Development of the Science Laboratory Assessment began in 1989 in response to a request for proposals from the California SDE/CTC. Several sources of information were utilized in developing the assessment materials and procedures to be pilot tested in the spring of 1990, including reviews of literature on effective science teaching, other teacher performance assessment systems, textbooks on science teaching methodology, California's curriculum guides and framework for science, and California's standards for beginning teachers. Over 100 California science educators were involved in the development of the assessment, either as members of the Assessment Development Committee or as reviewers of the assessment materials.

TABLE 4.5

DEVELOPMENTAL AND PILOT TEST COSTS FOR THE SCIENCE LABORATORY ASSESSMENT

Cost Categories	Development	Pilot Testing
Staff-Salaries & Benefits	$41,972	$ 9,869
Consultants (Teachers, assessors, and other consultants)	0	9,515
Travel (Consultants and staff)	0	3,664
Other Direct Costs (Site rental, phone, duplication)	7,790	2,348
Total Direct Costs	$49,762	$25,396
Indirect Costs	9,903	6,928
Total Costs	$59,665	$32,324

Although it was not possible for all Committee members to conduct tryout administrations of the assessment materials, several members did some type of activity related to tryout administrations. Only two committee members,however, completed a full assessment with a new teacher. Other members asked new teachers at their school to try out some part of the assessment (e.g., the questionnaire) or used the materials for self-assessment. As a result of these efforts, several modifications were made to the content and format of the materials.

Concurrent with the tryout administrations, a statewide review of the assessment materials was conducted by 63 science educators and scientists throughout California. Reviewers were asked to comment on several aspects of the assessment including the job necessity and appropriateness for new teachers of the domains and elements covered by the assessment. The developers reported that, overall, the reviewers seemed very positive about the materials and felt that the elements were necessary for effective teaching of a science laboratory activity and were appropriate to expect of a new teacher.

Reliability

The data reported in Figure 4.3 summarized the performance of the teacher candidates on this assessment. Since nearly all teachers passed most or all parts of the assessment, no further analysis was done to estimate the reliability of the assessment. The five instances in which two observers observed a teacher resulted in both observers rating the teacher as passing.

The pilot testing does not provide sufficient information upon which to judge the reliability of the assessment. It is not possible to determine at this point whether the teacher's performance reflects, (a) the degree to which all or nearly all teachers possess the skills reflected in the assessment, (b) the absence or unclarity of criteria for rating teachers which resulted in assessors assigning passing scores and being reluctant to assign a failing score in the absence of more definitive criteria, or (c) a need to build in greater range in rating which would allow assessors to better discriminate among teachers with differing levels of competence in those skills measured.

Two factors mentioned earlier can improve the scoring and information from this prototype assessment. First, more explicit criteria for scoring with supporting examples needs to be developed and incorporated into the training. This includes having assessors better summarize their observations by listing those factors which support and which indicate the deficiency of teachers on each of the factors. Second, consideration should be given to expanding the range of ratings so as to avoid the "ceiling" affect observed here which all teachers were rated similarly (i.e., as passing.)

Validity

The above section which describes development and background of this assessment provides information on the developers' involvement of science teachers and experts in the development of the prototype. This involvement contributed to the assessment's alignment with the curriculum frameworks and teaching standards which has been described. Thus, this information supports the content validity and focus of the assessment on important and current approaches to teaching science in laboratory settings.

Revisions to the scoring criteria and training may result in an assessment which better differentiates among teachers who are likely to have different degrees of skills in the areas examined by the prototype assessment. However, the pilot test only yielded information sufficient to say that the new teachers who participated in this pilot test were acceptable on the criteria as currently constituted.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Science Laboratory Assessment, organized into the areas of administration, content, format and a brief summary.

Administration of Assessment

As is often the case with high-inference observation instruments, the administration of the Science Laboratory Assessment is labor intensive, requiring nearly one professional person day per teacher. For this pilot test, each of the 11 experienced, science teachers who served as assessors agreed to conduct a minimum of three observations (i.e., take three days off from their teaching job or other work) during a six-week period. Few assessors were willing to leave their classrooms for more than the three days because of the difficulties they experienced trying to combine the administration of the assessments with the execution of their teaching duties. Therefore, should an observation system such as the Science Laboratory Assessment be considered for credentialing use in the state, we recommend the following:

- careful design of observation schedules for assessors to allow assessments to be distributed in a reasonable manner.

- consider expanding the recruitment pool of possible assessors to retired science teachers, science teacher supervisors, teacher trainers and others in addition to practicing teachers.

In addition to the above, the following factors seem to be key to smooth administration of the Science Laboratory Assessment in its present form:

- recruiting assessors who have expertise in more than one area of science (e.g., chemistry and physics) and/or experience teaching at or with different grade levels (e.g., high school and middle school) so as to allow more flexibility in the scheduling of observations;

- development of procedures for obtaining completed assessment materials from assessors in the field; and

- arrangements for storage of a large amount (at least 25 pages) of documentation per teacher.

Finally, since the Science Laboratory Assessment is administered and scored by the same person, the training of assessors is also a key factor to successful administration of the assessment. Through training, assessor candidates are taught the content of the assessment, as well as how to conduct and score the assessment. For this pilot test, training consisted of one home study day and two days of group training. However, based on assessors' comments, FWL staff's observation of the training, and FWL staff's review of the assessor's completed documentation forms, three days of training does not appear to be sufficient. Any future training might incorporate the following recommendations:

- increase group training time to no less than three days and possibly to five days;

- increase the training time allotted for introduction to, or review of, the assessment's content so that all of the participants agree on the definitions of the elements;

- include in the training more explicit instruction and examples on recording evidence and notes on each of the assessment forms, as well as on evaluating teacher performances; and

- increase the training time allotted for practice using the different forms to score teacher candidates and discussion of the results of this practice.

Following the above suggestions should greatly facilitate the administration of the assessment.

Assessment Content

Based on the observations of FWL staff, as well as information collected from assessors, teachers, and the assessment documentation (e.g., rating forms), the following conclusions are offered about the content of the Science Laboratory Assessment:

- Congruence of the Science Laboratory Assessment with the 1990 California Science Framework, Kindergarten through Grade Twelve, can best be described as partial. One way to strengthen the congruence would be to weave the idea of science themes--a major emphasis of the framework -- throughout more of the assessment (e.g., include as part of the elements and conference questions whenever possible).

- Coverage by the Science Laboratory Assessment of the California Standards for Beginning Teachers is relatively good. Coverage is particularly good for those standards which focus on student rapport and classroom environment, diverse and appropriate teaching, student motivation and conduct, presentation skills, and cognitive and affective outcomes of teaching. Coverage is partial for those standards addressing curricular and instructional planning skills, student diagnosis, achievement and evaluation, and a teacher's capacity to teach crossculturally.

- The job-relatedness of the Science Laboratory Assessment seems to be high because the assessment entails observing teachers actually teaching in their own classrooms.

- Overall, the content of the Science Laboratory Assessment does not seem too difficult for beginning teachers. Approximately 93% (27 of 29) of the pilot test participants received overall passing scores (i.e.,received an overall rating of "2"). Furthermore, at least 84% of the teachers passed each one of the assessment's seven domains.

- Analysis of the rating results by grade level (i.e., elementary, middle school, and high school) indicates that elementary teachers of science (i.e., those who have been trained to teach science at the elementary level) did as well or better on the assessments as did middle school and high school science teachers. Thus, the assessment seems an appropriate one for teachers of science at all grade levels.

- Acknowledging the increasing diversity in California's classrooms, the developers of the Science Laboratory Assessment included in the content of their assessment a domain specifically targeted to assessing a new teacher's ability to work with diverse students. This domain, however, was named by almost half the assessors as the hardest domain to rate because it depends as much on the assessor's knowledge of students in the classroom as on the teacher's knowledge. Thus, it is questionable whether the domain, as it is currently written, is an effective way of assessing a beginning teacher's ability to work with diverse students.

- The assessment was deemed by the teachers and assessors to be fair across groups of teachers (e.g., different ethnic groups, different language groups). However, as one teacher pointed out, true fairness depends on assessors' awareness of different teaching styles, especially with regard to management, climate, and communication. Without such awareness, the likelihood increases that the assessor may misinterpret or miss pertinent evidence.

- The majority of teachers and assessors think the Science Laboratory Assessment is an appropriate way of assessing (1) general teaching skills, and (2) skills in teaching laboratory science.

Assessment Format

One strength of the format of the Science Laboratory Assessment is that its focus is not on a simulated performance, or on how a teacher says s/he would perform, or on a teacher's knowledge of how to perform, but rather on a teacher's actual performance in the classroom. In addition, because the teacher is observed in his/her own classroom, no special facilities are required for administration.

Another strength of the format is that it actually includes two methods of assessment: observation and interview. The pre-and post-observation conferences which are part of the assessment are designed to (1) help the assessor understand the instructional goals and classroom context which affect the lesson design, and (2) give the teacher an opportunity to explain and justify changes in the original lesson design in response to unanticipated circumstances, as well as to reflect upon the lesson as it was conducted. The information provided in the two interviews and through the Pre-Observation Questionnaire (which is completed by the teacher before the observation) allows the assessor to conditionally evaluate teacher behaviors in light of differing instructional goals and classroom contexts. This type of observation instrument is superior to others used in teacher assessment because it focuses on the **meaning** rather than **frequency** of teacher behaviors.

Despite the above strengths, comments from the assessors and an analysis of the completed documentation and rating forms indicate that the format of the Science Laboratory Assessment could be improved in several ways. We suggest that consideration be given to following these recommendations:

- Shorten the **Pre-Observation Conference** either by eliminating or collapsing some of its 15 questions. Changes should be especially made which focus on improving the clarity and reducing the complexity of the questions.

- Because 82% (9 of 11) of the assessors said the **Guided Note-taking Form** (the form used by the assessors to simultaneously record and categorize the evidence/notes from the observation) was a source of difficulty, either training should be designed to specifically focus on the problems experienced by the assessors or the form should be greatly revised (and training should be designed to cover the revisions).

- The **Documentation Scoring Record** should be revised so that the two to four hours it takes to complete the form results in more than just a recopying of the evidence/notes from the various assessment components (e.g., observation, conferences). Perhaps evidence could be sorted not only by element, but also by whether it is positive or negative. Furthermore, not all evidence would have to be included, but only that which best exemplifies (positively and negatively) the elements.

- Although almost all of the assessors did not have difficulty rating the teachers' performances on the **Summary Report Form**, there was such little consistency in how the assessors wrote their summary remarks to support their ratings that this process needs substantial improvement. Any future training should include sufficient instruction as to how to write the summary remarks so that they (1) summarize the teacher's performance, and (2) adequately support/explain the assessor's rating.

- Because the Science Laboratory Assessment provides a wealth of information about the teacher's performance, the assessment's **rating process** should be revised so that this information is better utilized. The rating process might be revised from what is now essentially a pass/fail system to one which differentiates among teachers who are at different levels of performance. In this way, the instrument could also be used to increase teachers' competencies (e.g., through staff development) evidence/notes from the various assessment sources (e.g., observation, conferences).

Summary

If an observation system such as the Science Laboratory Assessment is selected as a method of assessing new teachers of science (or of other subjects) for credentialing purposes, the Science Laboratory Assessment could serve as a base upon which to build a fully developed prototype, but only after substantial revisions have been made to its documentation and rating processes.

CHAPTER 5:

LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT

The Language Arts Pedagogical Knowledge Assessment (LAPKA), developed by Northwest Regional Educational Laboratory, is a series of video-based exercises designed to assess the pedagogical content knowledge of elementary school teachers in language arts. The Spring 1990 pilot test version of LAPKA consists of four exercises, each of which is a videotape of a typical language arts classroom situation (i.e., scenario). The candidate's task is to view the videotape and respond in writing to a series of questions about the instruction depicted in the scenario. For some of the exercises, candidates also receive selected support materials (e.g., stories read by students) to assist them in their analysis.

Although each scenario depicts a language arts activity, they differ in the following respects: type of teaching activity, grade level, and group size. Scenario 1A and 1B, for example, are two versions of a teacher-led direct instruction lesson for a small group of first graders, while Scenario 2 depicts a teacher conducting writing conferences with individual students at different grade levels (i.e., 2nd, 4th, and 5th). Table 5.1 provides a summary of some of the characteristics of the four LAPKA scenarios. As described in the table, all of the scenarios depict a diverse student population (e.g., different ethnic groups, mixed abilities).

Each scenario is approximately 20 minutes in length; however, each exercise takes about an hour and 15 minutes to complete because of the time required to view the videotape (which is viewed in segments) and then write responses to the questions.

A more complete description of the content of each scenario follows:

Scenario 1A: Teacher-led, Small Group Reading Lesson. Scenario 1A is a teacher-led, direct instruction activity with a small group (eight students) of first graders. The general purpose of the lesson is preparing students to read. The focus of the activity is the story, Dragons and Giants, by Arnold Lobel. The videotape is divided into two sections. The first section features a pair of pre-reading activities: a vocabulary review and a word web. The second section shows the teacher and students orally reading the story, with particular attention given to the teacher's questioning strategies.

TABLE 5.1

SOME CHARACTERISTICS OF THE FOUR LAPKA SCENARIOS FOR ASSESSING A TEACHER'S PEDAGOGICAL CONTENT KNOWLEDGE IN LANGUAGE ARTS

	Scenarios 1A + 1B	Scenario 2	Scenario 3
Teaching Task	Teacher-led, direct instruction in reading	Student writing conferences	Literature discussion
Grade Level	1	2, 4, 5	4/5 combination
Group Size	Small group	Individual	Whole class and cooperative groups
Student Population	Range of reading abilities, ethnic backgrounds, and socioeconomic levels	Diverse abilities and ethnic backgrounds	Heterogeneous SES Students, predominately Hispanic, gifted program

Scenario 1B: Teacher-led Small Group Reading Lesson. Scenario 1B is another version of a teacher-led, direct instruction reading activity with a small group (ten students) of first graders. The purpose again is to prepare students to read. The teacher conducts the lesson around the story Cookies, by Arnold Lobel. She begins by introducing the concept of will power, using actual cookies as teaching props. After a group discussion (including brainstorming) about descriptive words, the group reads the story. Next, the students make paper cookies and graph their results. The lesson ends as the group prepares for poetry writing.

Scenario 2: Individual Writing Conferences. Scenario 2 consists of six writing conferences conducted by three different teachers. Each teacher works individually with a pair of students from either grades two, four, or five. The purpose for the writing conferences is to provide students with feedback on drafts of their writing. The conferences vary in length. A brief videotaped introduction precedes each pair of conferences.

Scenario 3: Cooperative Group Literature Discussion. Scenario 3 depicts a combination fourth/fifth grade gifted class in a cooperative learning activity in which the students examine the central theme of a story. Centered on a chapter from the novel, Sign of the Beaver, by Elizabeth Speare, the scenario is composed of three segments. The videotape begins with the teacher explaining her plans and goals for the lesson to the viewer. The second section shows the students working in small groups, reading orally, and discussing pre-formulated questions. The last section shows the teacher leading a discussion with the whole class.

For each scenario, the candidates are asked to respond in writing to three different types of questions: (1) description, (2) evaluation, and (3) extension. That is, the candidates are asked to describe important features of the pedagogical methods represented in the videotape, evaluate the effectiveness of these methods, and extend the principles underlying these methods to suggest ways of improving or changing the methods shown. Listed below is an example of each type of question:

EXAMPLES OF LAPKA QUESTIONS

DESCRIPTION: Describe at least three important ways the teacher prepared the children for reading this story. (Scenario 1A)

EVALUATION: Briefly comment on the strengths and/or weaknesses of this teacher's answer to the question on grading. (Scenario 2)

EXTENSION: In what ways might the lesson have been altered to increase its effectiveness? (Scenario 3)

The number of questions varies from scenario to scenario. Including subparts, the candidate responds to four questions in Scenario 1A, for example, but fourteen questions in Scenario 2. The candidate is expected to write about one paragraph per question, but the responses can be as brief as a phrase or as long as several paragraphs. Candidates are given approximately five minutes to respond to each question.

To score the performances, the candidates' written responses are compared against a list of possible acceptable responses or, in a few instances, general guidelines. For each response, the candidate receives either zero, one, or two points--with a maximum score possible for each question. The scores given at the question level are summed to give a total score for the scenario. While not specified in the developer's final report, it is likely that a candidate's scores for all the scenarios would then be totalled and reported as a single, cumulative score.

Administration of Assessments

The following section presents a discussion of the logistics of administering the assessment, security issues, requirements for test administrators, and teacher and FWL impressions of administration.

Overview

LAPKA was administered at three sites, one in the Bay Area and two in Southern California between April 21 and May 12, 1990. As seen in Table 5.2, 42 beginning (first- or second-year) teachers participated in the pilot test. Four of the teachers were males and ten were members of minority groups. There were 21 teachers at grades K-3 and 21 at grades 4-7. Slightly more than three-quarters of the teachers completed their teacher education preparation in California higher educational institutions, and one quarter received their preparation outside the state of California. Twenty-five teachers had taken two or more methods courses in reading; seventeen had taken none or one. All of the teachers except one indicated that some of their students spoke languages other than English.

TABLE 5.2

PILOT TEST PARTICIPANTS

LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT (LAPKA)

(Number of Teachers = 42)

Descriptive Characteristics of Participants	Number
Gender	
Male	4
Female	38
Ethnicity	
Asian	1
Black	5
Hispanic	4
Native American	0
White	32
Grade Level	
K-3	21
4-7	21
Teacher Training Program	
In California	33
Outside of California	8
No Response	1
Number of Reading Methods Courses	
0-1	17
2 or more	25

Each pilot test session was conducted in a four-and-one-half hour session by one or two test administrators. Each administration included an overview, two fifteen-minute breaks, and three of the four exercises (called scenarios). The teachers at the first two pilot test sessions completed Scenarios 1A, 2, and 3; the teachers at the third session completed Scenarios 1B, 2, and 3. (It should be noted again that 1A and 1B were two versions of the same type of exercise.)

Logistics

Administration required the following logistical activities: identifying a sample of teachers, sending orientation materials to teachers, administering the assessment, and acquiring feedback from the teachers.

Identifying teacher samples. In recruiting elementary school teachers for the LAPKA pilot test, FWL aimed to identify a diverse group of teachers from a variety of settings. At the same time, for the pilot test administration to be logistically feasible and cost effective, we needed to identify administration sites that could be reached by car or public transportation by a sufficiently large number of beginning teachers in an hour or less. With these conditions in mind, we contacted a number of project directors from the California New Teacher Project in Southern California and the Bay Area. These project directors supplied FWL with a list of names and school sites of first- and second-year teachers in their project. FWL contacted these teachers by phone to ask for their participation. All teachers were offered $80 to participate in the pilot test.

Sending orientation materials. The orientation materials sent to teachers for this assessment included a two-page overview which described the content and format of the assessment, a brief description of the California New Teacher Project and its Assessment component, and directions to the administration site.

Assessment administration. Although no special training is needed for the administrator(s) of this assessment, the role of the administrator(s) in this pilot test was a key one requiring the following administrative activities: First, test materials were distributed at the beginning of the session in three separate manila envelopes, one envelope for each scenario. Each envelope contained the following materials: (a) instructions to viewers, (b) support materials (e.g., story or chapter of a book), and (c) several sheets of questions. The test administrator then instructed the candidates to remove and read the sheet(s) of questions pertaining to the video segment about to be viewed. After the candidates had a chance to read the questions, the administrator showed the segment. The administrator then turned off the videotape and instructed the candidates to answer the questions pertaining to the segment. When the allotted time for answering the questions expired, the administrator instructed the candidates to read the next sheet(s) of questions pertaining to the next video segment. The administrator then showed the next video segment and repeated the process described above. Each of the three video exercises was administered in this manner.

The three separate administration sessions were conducted in average-sized classrooms or staff rooms with groups ranging in size from twelve to twenty-five. For the two smaller groups, a single video monitor was adequate; for the larger group, two video monitors were necessary to insure that all of the candidates could clearly see and hear the tape. LAPKA could potentially be held in a large conference room with several hundred candidates taking the assessment at the same time, or be carried over closed-circuit television and shown in a number of rooms at the same time. The only requirements would be to have a video monitor for approximately every twenty candidates and enough staff on site to monitor candidates during the assessment.

Collecting evaluation feedback. Immediately after viewing the videotapes and answering the questions, the teachers were asked to complete an evaluation feedback form in which they gave their thoughts and opinions about the assessment.

Security

For security purposes, the format and focus of LAPKA, as well as many of the actual questions, could remain unchanged with each new administration, but there would need to be a change in the content of the assessment each time. Also, while the overall scoring system could remain the same, new scoring criteria would need to be developed for any changes in the content. Thus, some development costs for LAPKA would be ongoing, but these costs might be significantly diminished on a per candidate basis if the assessment is administered simultaneously to a number of large groups of candidates.

Another security-related issue is the influence of coaching. The coaching that might take place, however, would likely contribute to the teachers' professional development. For example, two typical questions in LAPKA ask the teacher to (1) identify two effective features of the teacher's language arts instruction, and (2) identify one important way that the teacher's instruction could have been enhanced. To practice for this assessment, teachers might join together to view videotapes of each other's language arts lessons and discuss them using the above two questions as focal points. This kind of "coaching" would likely improve not only teachers' performance on LAPKA, but their language arts practices as well. To prepare for this assessment, candidates might also memorize a "generic" list of effective language arts practices (e.g., activate students' background knowledge before reading), but learning such a list without developing an understanding of the principles from which the items on the list are derived is not likely to significantly improve teachers' scores on LAPKA (or their actual teaching). Security would be addressed through the use of new video taped segments and scoring criteria for them.

Scoring

The following section includes an explanation of the scoring process, a description of the training of scorers, and a discussion of scorers' perceptions, as well as those of FWL staff, of the training.

Scoring Process

As mentioned earlier, the candidates' responses to the assessment questions are scored by comparing them to the responses listed on a scoring key provided by the assessment developer. The scoring key lists both acceptable and unacceptable responses for each question, and also stipulates the number of points (i.e., zero, one or two) to be awarded for each response, as well as the maximum allowable number of points for any single question.

The number of possible acceptable responses described in the scoring key ranges from two to ten for each question. For example, in the scoring key for Scenario 1A/Question 1, the candidate is asked to "describe at least three important ways the teacher prepared the children for reading this story." The scoring key lists five acceptable responses for this question, allows one point for each correct answer, and sets the maximum number of points to be awarded at three. The acceptable responses listed in the scoring key for most of the questions are intended by the developer to be exhaustive; that is, any response by a candidate that does not correspond to one of the items on the scoring key is not awarded credit.

The scoring key also provides examples of responses that should not be credited. In Scenario 1A/Question 1, three "no credit" responses are listed. For example, no points are awarded to a candidate who points out that the teacher "uses the table of contents to locate the story." The listed unacceptable responses are intended only as examples, but the most commonly expected responses that would not receive credit are included in the key.

For a few questions, the scoring key presents broad guidelines rather than specific examples of acceptable responses. In these cases, the scorer is expected to rely more heavily on his/her professional judgement, than on the scoring key. The scoring manual also does not provide an explicit rationale why some responses are acceptable and others are not, nor why some responses are awarded one point and others two points.

The teachers' responses to the assessment questions were all scored during the two-day training period for the scorers. This training session is discussed below.

Scorers and Their Training

The training of scorers and the scoring of the LAPKA pilot test responses took place at FWL over a two-day period, on June 26 and 27, 1990. The training and scoring session was directed by two staff members from NWREL, who were also members of the LAPKA development team. Six scorers participated in the two-day session.

Scorer characteristics. The six scorers were all current or former teachers. Four were practicing elementary school teachers, ranging in experience from six to thirty years. One of the scorers was a former elementary school teacher and principal, and one was a former secondary English teacher who was a graduate research assistant with Stanford University's Teacher Assessment Project in elementary language arts instruction. All six scorers were female; one was African-American, and another was Asian-American.

Training. The two-day training session was roughly divided into four half-day sessions, with each session devoted to the training of scorers and the actual scoring of candidate responses for one of the four scenarios. The trainers opened with a brief overview of LAPKA and then proceeded with training and scoring each scenario in order from Scenario 1A through Scenario 3.

The procedure for training was as follows: One of the trainers began by presenting a brief overview of the scenario. The scorers then read relevant materials, such as the literature selection and instructions to the viewers, previewed the assessment questions for the upcoming segment of the videotape, and viewed a segment of the scenario. The trainers explained the scoring key for that segment of the exercise, then the six scorers individually scored the same two candidate responses for that segment and discussed any scoring-related issues. After the segment-by-segment training session for the scenario was completed, pairs of scorers independently scored sets of candidate responses for that scenario.

Because Scenarios 1A and 1B were each completed by only half of the pilot test candidates--1A by one half of the candidates, and 1B by the other half--they required slightly less time to score than Scenarios 2 and 3. Each scenario, with the exception of Scenario 3, was double scored. Scenario 3 was scored by a single scorer because of time constraints. Thus, interrater reliability scores were only available for Scenarios 1A, 1B, and 2. Each pair of scorers was assigned a subset (one-third) of candidate responses to score. The two scorers in each pair both scored the same candidate responses but did so independently. The three scorer pairs were reassigned so that each scorer was paired with a different partner for each scenario.

To help facilitate the scoring process, the trainers provided scoring sheets for the scorers to use in marking their scores. These sheets listed the individual scoring criteria down the right side of the page and blank lines on the left side of the page to indicate whether the candidate received credit for a particular criterion (see Appendix C for an example of a scoring sheet).

Perceptions of training. The scorers' perceptions of the training they received was mixed: two rated the training as very good, two as adequate, and two as poor. The scorers offered several suggestions for improving the training. One of their major concerns was that the training was too brief. They thought that they were asked to score the pilot test responses before they had adequately reached a shared understanding of (1) the scoring criteria and (2) how narrowly or broadly to apply the criteria to candidate responses. The low interrater reliability scores (discussed later in the Technical Quality section) suggest that scorers were not interpreting the criteria in a consistent fashion. The scorers recommended that more time be allotted for training, particularly for working with practice materials.

In addition, rather than viewing the scenarios segment-by-segment, the scorers thought it would have been more valuable to first watch each scenario from beginning to end without interruptions before attempting to score any candidate responses. This procedure, they said, would have helped them develop a better overall perspective for scoring the exercise.

Finally, the scorers also felt that they were being asked to narrowly and mechanistically apply the scoring criteria, when they should have been given more latitude to interpret candidate responses and apply their professional judgment. This issue was very important to the scorers and is discussed more extensively in the format section of this report.

FWL staff agree with the scorers' recommendations that (1) the training be lengthened, and (2) the training should provide scorers with an overall understanding of the entire scenario and accompanying questions before any scoring is attempted. Other comments pertinent to any future training sessions can be found in the format section.

Teacher and FWL Staff Impressions of Administration

FWL staff members administered LAPKA on three separate occasions. No significant problems arose at any of the sessions. While a few of the students' comments on the videotape were inaudible (from any distance), none of the teachers reported that this problem interfered with his or her ability to respond to the exercise questions. Many of the teachers did, however, report fatigue from the four-and-a-half hours of assessment.

Assessment Content

The content of the LAPKA scenarios focuses on assessing a teacher's pedagogical content knowledge in the following areas:

(a) reading;

(b) writing; and

(c) response to literature.

As described earlier, the assessment asks candidates to view videotaped segments of language arts lessons which focus on each of the above areas, and then to comment on the pedagogical practices which are depicted. After viewing each scenario, the candidates are asked a variety of questions to elicit their knowledge about (1) the pedagogical content method(s) used by the teacher in the videotape, (2) the rationale for, or the effectiveness of, these method(s), and (3) other pedagogical content methods which could be used instead of or in addition to those depicted.

While the LAPKA scenarios do not cover the entire range of instructional activities carried out by teachers in language arts, they represent a diverse set of activities that are essential to any successful language arts program. In addition, these scenarios cover instructional situations across a broad span of grades and student ability levels. It should be noted, however, that while the three areas listed above are generally acknowledged to be major components of a language arts program, the developers did not intend for the content of the assessment to provide a representative assessment in language arts. Instead, they chose these three areas as a focus for an alternative assessment approach--i.e., the use of videotapes as a stimulus for written responses.

In the following pages, the content of LAPKA is evaluated along the following dimensions:

- Congruence with the California English/Language Arts Model Curriculum Guide for Kindergarten through Grade Eight;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the instrument;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, diverse student groups);
- Fairness across groups of teachers (e.g., ethnic groups, gender);
- Appropriateness as a method of assessment.

Congruence with California Curriculum Guides and Frameworks

FWL staff reviewed the content of the LAPKA assessment to see in what ways it is congruent with California's English-Language Arts Model Curriculum Guide for Kindergarten through Grade Eight (SDE, 1987).

The guide presents recommendations for an effective English-language arts program in the form of twenty-two statements, referred to as guidelines. These guidelines are categorized in five major groupings: (1) the reading and studying of significant literary works, (2) classroom instruction based on students' experiences, (3) integration of the language arts, (4) integration of English-language arts with other subject matter areas and settings outside the classroom, and (5) student evaluation. Within each of the five major groupings are several guidelines which focus on specific features of an effective language arts program. The following section examines the congruence of LAPKA with these general groupings and guidelines.

Grouping 1: The English-language arts program emphasizes the reading and the study of significant literary works. The three guidelines within this grouping stress the importance of providing intensive, direct instruction for all students in comprehending and responding to core works of literature, creating opportunities for students to explore and extend their experiences with literature, and supporting and encouraging students to read independently.

Literature is emphasized in three of the four LAPKA scenarios (i.e., 1A, 1B, and 3). Instruction in these scenarios revolves around a literary work such as a children's picture book by a well-known author or a children's historical novel which is on the state's list of recommended literature (SDE, 1986). The candidate's task is to describe and evaluate the teachers' approach to instructing the students in comprehending and responding to the works of literature. The candidates are also asked to offer suggestions of other ways to provide such instruction. Thus, the emphasis on literature in LAPKA is strongly congruent with the curriculum guide.

Grouping 2: The English-language arts program includes classroom instruction based on students' experiences. The two guidelines within this grouping deal with the importance of having students draw on their experiences while participating in language arts activities, and of students participating in activities designed to give them experience and knowledge needed to be proficient in the language arts.

The teachers in the videotaped scenarios are shown providing instruction that draws upon their students' background knowledge and experiences. The candidates being assessed are expected to recognize this instructional practice as well as to suggest alternatives or ways to improve upon what the teachers actually did. The focus in LAPKA on basing instruction on students' knowledge and experiences is strongly congruent with the standards in this grouping.

Grouping 3: English-language arts instruction is based on an interrelated program in which listening, speaking, reading, and writing, with literature at the core, are taught in concert and are mutually reinforcing. The eight guidelines in this grouping focus on the development and integration of speaking, listening, reading, and writing skills and strategies. This group of guidelines also addresses the teaching of decoding strategies, the conventions of the English language (e.g., spelling, punctuation), and handwriting skills.

In the four videotaped LAPKA scenarios, the teachers engage their students in a variety of reading, writing, speaking, and listening activities. Candidates being assessed are asked to describe and evaluate the teachers' instructional practices and to offer alternative ways that they might approach similar teaching tasks. LAPKA's focus on the development and integration of the language arts in these scenarios is congruent with most of the guidelines in this grouping. LAPKA does not, however, address the teaching of decoding strategies, language conventions, or handwriting skills.

Grouping 4: English-language arts are an integral part of the entire curriculum. The seven guidelines in this grouping focus on the connection between English-language arts skills and other subject matter areas. Some of the topics covered are the use of higher-order thinking skills in English-language arts and other subject areas, broadening students' vocabulary, using the library and other media and technological resources, modeling of communication skills by school staff members, and involvement of parents in the educational program.

Most of these guidelines fall outside of the focus of LAPKA. LAPKA does not, for instance, deal with the integration of the language arts in other subject areas or outside the classroom. Scenario 1B does, however, present the teacher weaving another content area (e.g., math) into the reading lesson. For Scenarios 1A and 1B, LAPKA also asks candidates to identify important features of the videotaped teachers' vocabulary instruction, and for Scenario 3, candidates are asked to discuss the teacher's goal of encouraging students to think in new directions (i.e., higher-order thinking). Thus, congruence of LAPKA with this grouping should perhaps best be described as partial.

Grouping 5: Evaluation of the English-language arts program includes a broad range of assessment methods. The two guidelines in this grouping stress the importance of multidimensional measures of assessment and student self-assessment. Scenario 2 addresses teachers' evaluations of student writing with some self-assessment by students included; the other scenarios do not address teacher assessment of students, student self-assessment, or program assessment. Congruence with this grouping therefore is weak.

Table 5.3 summarizes the congruence of LAPKA with the *English-Language Arts Model Curriculum Guide for Kindergarten through Grade Eight*. Overall, LAPKA is strongly congruent with the guide in some areas and partially congruent in others.

TABLE 5.3

CONGRUENCE OF THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT (LAPKA) WITH THE ENGLISH-LANGUAGE ARTS MODEL CURRICULUM GUIDE FOR KINDERGARTEN THROUGH GRADE EIGHT

Model Curriculum Guide Content	Relevant LAPKA Content	Extent of Congruence
Grouping 1: The English-language arts program emphasizes the reading and study of significant literary works.	-Study of literature is central to 3 of 4 scenarios.	Strong
Grouping 2: The English-language arts program includes classroom instruction based on students' experiences.	-Role of student background knowledge is emphasized in all scenarios.	Strong
Grouping 3: English-language arts instruction is based on an interrelated program in which listening, speaking, reading, and writing, with literature at the core, are taught in concert and are mutually reinforcing.	-Instruction portrayed in the scenarios is based on integration of language arts. The scenarios do not address decoding strategies, language conventions, or handwriting skills.	Partial
Grouping 4: English-language arts are an integral part of the entire curriculum.	-The scenarios do not address language arts instruction outside of the classroom. Scenario 1B addresses integration of language arts with other subject matter areas. Scenario 3 addresses critical thinking. Scenarios 1A and 1B address vocabulary development.	Partial
Grouping 5: Evaluation of the English-language arts program includes a broad range of assessment methods.	-Scenario 2 addresses teacher evaluation of student writing. The other scenarios do not address evaluation.	Weak

Extent of Coverage of California Standards for Beginning Teachers

The California Beginning Teacher Standards are criteria for teacher competence and performance that the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. Listed below are brief italicized descriptions of the Standards (22 - 32) that pertain to expectations of student competencies to be attained prior to graduation from teacher preparation programs. (Standards 1 through 21 address programmatic requirements.) To evaluate this assessment instrument and make inferences about the assessment approach which it represents in terms of its appropriateness for use with California elementary teachers, the stimulus materials and scoring criteria for each exercise were compared with the eleven relevant California Beginning Teacher Standards. Each standard is discussed separately.

Given that these standards are intended to guide the evaluation of teachers' performance in the classroom, an assessment such as LAPKA only indirectly addresses these standards. LAPKA measures teacher knowledge, but it does not provide direct evidence of teachers' ability to translate that knowledge into actual practice.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* LAPKA does not address this standard. The teachers in the videotapes vary in their approaches to creating a positive and productive classroom environment, but the candidates are not asked to comment on this feature of instruction.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* While the candidates in this assessment do not plan lessons of their own, this standard is addressed in a limited fashion in LAPKA. In Scenarios 1A, 1B, and 3, candidates are asked to evaluate the features of another teacher's lesson and suggest alternatives and extensions to these lessons, which provides some indirect evidence of the quality of lessons that candidates might be capable of designing and carrying out in their own classrooms.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* LAPKA does not provide even indirect evidence of a teacher's abilities in this area. While a variety of instructional strategies are portrayed on the videotapes, candidates are not asked to comment on the appropriateness of these various strategies for meeting the diverse needs, interests, and learning styles of students.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* LAPKA addresses this standard in a limited way. One of the candidate's tasks is to identify effective features of the various instructional activities portrayed in the four different scenarios. The candidate's comments on the ways that the videotaped teacher successfully motivates students, such as through pre-reading activities (e.g., Scenario 1A/Question 1) or suggestions for motivating students, such as through pre-writing activities (e.g., Scenario 2/Question 1B), could provide some indirect evidence of a candidate's ability to motivate and sustain student interest and participation.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* This standard, which focuses on teachers' presentation and communication skills in the classroom, is not addressed by LAPKA.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* LAPKA does not provide evidence about a candidate's ability to achieve his or her instructional objectives or to assess the achievements of students in a class.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* This standard is addressed to a limited degree by Scenario 3. The videotaped teacher presents a literature-based lesson in which one of her goals is to develop students' decision-making and critical thinking abilities. The candidate's task for several of the questions is to discuss why it is an important goal for a language arts activity. Thus, a candidate's response to these questions provides some evidence of his or her knowledge of ways to improve the thinking ability of students.

Standards 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* A few questions in LAPKA address this standard in a limited way. For example, in one of the individual writing conferences in Scenario 2, a teacher criticizes a student's writing at great length, and the student begins to visibly retreat by folding his arms and moving back. A candidate's comments about this encounter might provide some information about that his or her sensitivity to the affective dimension of teaching.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socioeconomic differences.* LAPKA does not address this standard. While the students portrayed in these scenarios come from a range of socioeconomic, linguistic, and cultural backgrounds, the assessment does not provide any evidence about a candidate's capacity to teach students who are different from the candidate.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* This standard is partially addressed in LAPKA. The four videotaped scenarios in LAPKA cover the span from elementary to middle school ages, and include a range of individual and group ability levels. The teachers in these scenarios present diverse instructional strategies for promoting the reading, writing, and oral language development of students of various grades and abilities. Scenarios 1A and 1B present a direct instruction lesson in reading for first and second graders; Scenario 2 focuses on individual writing conferences across several grade levels; and Scenario 3 portrays a literature-based cooperative group activity in a fourth/fifth grade class. A candidate's comments about the effective and ineffective features of the instruction shown in these scenarios, along with his or her suggested extension activities, offers some indirect evidence of a candidate's potential to provide instruction that takes into account the needs of students of different ages and abilities.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interaction with other members of the profession.* This standard is not addressed in LAPKA.

The extent of coverage by LAPKA of the California Beginning Teacher Standards is summarized in Table 5.4. The table lists the LAPKA scenarios that address each standard, and also describes the extent of coverage provided.

Job-Relatedness

The teacher candidates who took the assessment and the scorers who evaluated the candidates' responses strongly agreed (89%) that the pedagogical content knowledge assessed by LAPKA is relevant to the job of teaching elementary language arts. However, several of these teachers questioned the reality of reduced class sizes and individual writing conferences portrayed in the videotapes:

> *The teacher-led instruction had only ten kids. Get real!*
>
> *No teacher has time for one-on-one writing instruction.*
>
> *It is an unrealistic scenario [i.e., individual writing conferences] for a teacher responsible for 30+ kids.*

FWL staff agrees that the pedagogical content knowledge assessed by LAPKA is related to the job of teaching elementary language arts. As for the teachers' comments about some of the scenarios not being relevant to a "real" teacher's job, FWL acknowledges that extended individual writing conferences, for example, are probably difficult to conduct in most teachers' classrooms; however, the areas of pedagogical content knowledge (e.g., writing, reading) assessed by the scenarios are relevant to any job of an elementary teacher of language arts, and are not strictly related to the size of the groups depicted in the scenarios.

Appropriateness for Beginning Teachers

The discussion in this section focuses on the teachers' perceptions of the appropriateness of LAPKA for beginning teachers and their performance on the assessment.

Perceptions. Most (78%) of the candidates reported that LAPKA was of appropriate complexity for assessing new teachers and that they have had sufficient opportunity to acquire the knowledge and skills necessary to respond to the assessment questions.

TABLE 5.4

EXTENT OF COVERAGE BY THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT (LAPKA) OF THE CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	LAPKA Scenarios Addressing Standards	Extent of Coverage
22: Student Rapport and Classroom Environment	–None	None
23: Curricular and Instructional Planning Skills	–Scenarios 1A, 1 B, and 3	Partial
24: Diverse and Appropriate Teaching	–None	None
25: Student Motivation, Involvement and Conduct	–Scenarios 1A, 1B, 2 and 3	Limited
26: Presentation Skills	–None	None
27: Student Diagnosis, Achievement and Evaluation	–None	None
28: Cognitive Outcomes of Teaching	–Scenario 3	Partial
29: Affective Outcomes of Teaching	–Scenarios 2 and 3	Limited
30: Capacity to Teach Crossculturally	–None	None
31: Readiness for Diverse Responsibilities	–Scenarios 1A, 1B, 2 and 3	Partial
32: Professional Obligations	–None	None

While the majority of the teachers felt prepared to take this assessment, several of them emphasized that most of what LAPKA taps is learned on the job after they have completed their education courses and student teaching:

> *I would have not been able to answer these questions just based on coursework and student teaching. Much of what was assessed was learned during my first year, not before.*

> *I think you'd need to teach before taking this. Books or class knowledge doesn't necessarily give you the answers.*

> *I think all [of this assessment] should be given after experiencing a real classroom on your own. It's very different than student teaching.*

One teacher felt that a lack of experience accounted for her difficulty in answering some of the "extension" questions:

> *I felt slightly inadequate to continue coming up with enhancement ideas. It seems that experience plays an important role in creating a variety of ideas and methods to be used in instruction.*

A few teachers reported that they did not feel prepared for almost any of the assessment:

> *A lot of things I have never learned, either in the education program or in the new teacher program. My language arts in college taught me hardly any of this.*

Other teachers felt unprepared for certain sections of the assessment, particularly those that dealt with a grade level they had not taught or a subject area (e.g., writing instruction) with which they had limited experience.

While the developers of LAPKA initially planned to portray only good teaching practices, the scenarios that were produced did not always achieve this goal. In response to those scenarios that did depict exemplary teaching, however, many of the teachers, in written and oral feedback, expressed excitement about the teaching practices portrayed, as well as an eagerness to try out similar methods in their own classrooms. One teacher commented:

> *I thought [the assessment] was very enjoyable. I think teachers ought to be able to see other teachers teaching model lessons in order to gain insights into their own teaching strategies.*

Given that this assessment is aimed at beginning teachers, consideration should be given to the question of whether less-than-examplary teaching practices should be presented. On the one hand, as indicated by the teachers' comments above, scenarios that depict exemplary teaching practices offer the instructional benefit of affording teachers the opportunity to learn how to improve their teaching. On the other hand, one strength of videotapes is their capacity to **show** those things--both positive and negative--that a paper and pencil assessment can only describe (i.e., a picture is worth a thousand words). For example, LAPKA's scenario of a young student becoming very discouraged during a writing conference with his teacher demonstrates in a much clearer and more powerful way that a student's self-esteem can be immensely affected by a teacher than could ever be demonstrated in a short written description. Thus, scenarios of less-than-exemplary teaching practices could be used to assess a candidate's knowledge of incorrect or undesirable teaching practices. There is the risk, however, that the teaching practices portrayed would not be recognized as undesirable, simply because they are part of a state assessment. The questions accompanying such a scenario would have to be carefully designed so as to avoid this possibility.

Performance on assessment. FWL staff analyzed the teachers' overall performances as well as their scores on each of the four individual scenarios to see if the beginning teachers participating in this assessment had acquired the knowledge and skills measured by LAPKA. For Scenarios 1A, 1B, and 2, candidate scores represent the average of two independent ratings, while the scores for Scenario 3 are the ratings given by a single scorer.

The total number of points possible, the means, standard deviations, and ranges for each individual scenario are reported on Table 5.5. The candidates' average scores on the scenarios ranged from 57% to 87% correct, suggesting that the assessment is probably of appropriate complexity for beginning teachers. Making a few changes in the assessment format, such as eliminating or rewording the few ambiguous questions and more clearly marking the spaces for candidate responses, particularly in Scenario 2, would possibly result in higher scores.

The candidates had the most difficulty with Scenario 2, the individual writing conferences, correctly answering only 57% of the questions. The candidates scored the

TABLE 5.5

PERFORMANCE DATA FOR PILOT TEST TEACHERS (N=42) FOR THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT (LAPKA)

Scenario	Total Possible Points	Mean	Percent Correct	Standard Deviation	Range
Scenario 1A	13	8.80	68%	1.97	2–11
Scenario 1B	4	3.47	87%	0.66	2–4
Scenario 2	39	23.38	57%	6.03	11–37
Scenario 3	35	24.14	69%	5.00	12–33

highest on Scenario 1B, a direct instruction reading lesson, getting 87% correct. However, on Scenario 1A, which is another version of a direct instruction reading lesson, the candidates scored only 68%. While the difference in the average candidate scores for Scenarios 1A and 1B suggests that these two scenarios may have been tapping different areas of candidate knowledge, the small number of questions in Scenario 1B (4 points) makes any comparison between these scenarios tenuous.

There was a wide range of scores for each scenario. For each scenario, several candidates obtained near perfect scores, (or, in the case of Scenario 1B, perfect scores), while other candidates missed a large percentage of the questions. This outcome suggests that LAPKA may effectively discriminate between weaker and stronger candidates.

The performance data indicated no notable differences in candidate performance based on gender, ethnicity, grade level taught, California or non-California teacher training program, or the number of reading methods courses taken.

Appropriateness across Contexts

The LAPKA scenarios portrayed language arts instruction in a variety of contexts--across grade levels, with a diverse group of students, across the language arts, and with a variety of instructional settings. The following sections look at the perceptions of the teachers regarding the appropriateness of the assessment across certain contexts, as well as the perceptions of Sharon Nelson-Barber, our consultant on cultural diversity.

Grade level. The multiple subjects credential issued to elementary teachers spans the grades K-8. Thus, any assessment for an elementary teacher should in some way address that teacher's capacity to teach the grade levels covered by the credential. The LAPKA assessment includes scenarios depicting instruction by teachers to students of different grade levels, and assesses candidates on their knowledge about this instruction. Most of the teachers (68%) who participated in the pilot test of LAPKA believe that the assessment is appropriate for teachers of different grade levels. One teacher commented:

> *You never know what grade level you'll be teaching--therefore, [you] need to be accountable for all.*

Diverse students. Most of the teacher candidates (63%) indicated that this assessment is appropriate for teachers of diverse student groups. Several of the teachers who disagreed gave the following comments:

> *I saw very few if any blacks or Hispanics. The [videotape] didn't show California ethnic groups.*
>
> *There could be some videos with LES or bilingual classroom settings.*
>
> *[The videotape didn't show] classrooms where students are very disruptive or where reading ability [is] very low.*

The above comments are important ones, especially in light of the fact that one of the advantages of using videotapes is the ability to **show** a mixture of students in a classroom--not just describe the students. Moreover, those students who usually present a challenge to beginning teachers are those of limited-English proficiency or low reading ability--and not the typical GATE students as shown in Scenario 3. Should the LAPKA assessment be developed further, consideration might be given to including scenarios that depict students whose diversity is not only commonly represented in California classrooms, but also commonly presents a challenge to beginning teachers.

Sharon Nelson-Barber, our consultant on cultural diversity, brings up another issue to be considered with regard to LAPKA's appropriateness across contexts. It is her belief that the LAPKA assessment is built around a conceptual framework that consists of "certain assumptions about how classrooms should be organized, how students should be rewarded, how talk should proceed, etc." These assumptions, according to Nelson-Barber, have "the potential to miss many of the instructional techniques and interactive behaviors deemed effective in some minority communities." As she comments,

> *Teaching in multicultural classrooms requires going beyond the teaching of content to the relationship of that content to students' broader contexts--their social environments, their communities, their attitudes, even their feelings.*

For example, teachers of black students in predominantly black communities may respond to that environment by viewing the teaching of basic skills as essential to their students' survival in the mainstream community. They may put a focus on grammar, punctuation, spelling, etc. as a way of moving their students towards mastery of the mainstream language. Similarly, teachers of classes of predominantly Asian students in predominantly Asian communities may respond to the high academic expectations of the community by focusing on grading or by considering grading to be an important part of their classroom gestalt.

According to the LAPKA assessment, however, it is likely that teachers in the above contexts would be penalized. For Scenario 2, Part A, Question 2, for example, if the candidate makes reference in his/her response to correcting grammatical, punctuation, spelling errors, etc., the candidate is to be awarded zero points. For Question 1, Part D, of the same scenario, the candidate is also awarded zero points if s/he emphasizes the importance of grading without making reference to the value of composing or organizing thinking. Although the LAPKA scenarios do not depict classes or groups largely composed of black students or Asian students, teachers who are used to such contexts may have a concept of teaching which differs from that of the assessment developers.

Another example given by Nelson-Barber is that of teachers of students "whose only prior interaction with text may have been holding a hymnal or observing the priest reading in a foreign tongue." For these students, preparation for reading a story may necessitate such basic preliminaries as asking the students to locate the story in the table of contents. The LAPKA assessment, however, does not recognize such preparation; if a teacher mentions directing students to the table of contents in his/her response to Question 1, Scenario 1A, zero points are awarded for the answer.

It might be argued that, because the LAPKA scenarios do not show classrooms or groups of students as described by Nelson-Barber, none of the teachers' responses should resemble those suggested by Nelson-Barber. This argument, however, does not take into consideration the beginning teacher's experience and general approach to teaching. Teachers who are used to working with black students in predominantly black communities, or with Asian students in predominantly Asian communities are likely to respond to the LAPKA questions with those contexts in mind. Although the answers they give may not be appropriate for the students in the videotape, should they be penalized for describing practices that research claims are effective for the students they teach?

FWL staff suggests that consideration be given to incorporating into the assessment a way of ascertaining the teacher's philosophy of instruction and context of experience so that his/her responses to the assessment questions can be judged within that framework. Without such a revision to the assessment, it is difficult to conclude that LAPKA is fair across contexts.

Fairness across Groups of Teachers

While a high percentage (89%) of the teacher candidates felt that this assessment was fair to new teachers of both genders, different ethnic groups, and different language groups, this view was not shared by some of the teachers, the scorers, and our consultant on cultural diversity. Of major concern was the lack of teacher diversity shown on the videotapes. One of the scorers, an Asian-American female, summed up the problem as follows:

> *Future LAPKA assessments need to include videotapes that show men as well as Hispanic and African-American teachers. Although the majority of teachers in California currently are Caucasian, given changes in the state's demography and the need for better minority representation in the teaching pool, assessment exercises used for certification should show the ethnic diversity of teachers that are found in the state's schools.*

Nelson-Barber, our consultant on cultural diversity, agreed that the videotapes should have presented more diversity with respect to the teachers portrayed giving the lessons, but she also warned that presenting a diverse group of teachers will serve little purpose if the "diverse group of teachers demonstrates only one view of teaching."

As was described in the section above, distinctly different approaches to teaching are often utilized by teachers of different student groups. Nelson-Barber points out that the teachers most likely to use these different approaches are teachers of the same racial or ethnic group as their students. Black teachers of black students, for example, are very likely to emphasize grammar, spelling, punctuation, etc. in their instruction (Delpit, 1986, 1988)--a teaching practice penalized by the LAPKA assessment (see the section above). Research by Kleinfeld (1974) shows that the degree of "teacher-directedness" or the proximity of teacher to students can make real differences in the educational lives of American Indian students, who are accustomed to the "affectively intense and particularistic relationships characteristic of small traditional societies." Native teachers of American Indian students are most likely to view management and instruction as intimately tied--another practice penalized by the LAPKA assessment (e.g., the script for Scenarios 1A and 1B explicitly asks candidates to ignore the videotaped teacher's management practices; candidates are awarded zero points if they mention management issues or management-related activities in their answers to questions in Scenarios 1A and 2).

Still another example supplied by Nelson-Barber is that of black teachers displaying a great deal of emotion with black students. Although an outside observer (or an uninformed assessor) may perceive such a teacher as "authoritarian, pushy, or harsh," Nelson-Barber cites research that shows that "for some members of the African-American communities, teachers who do not exhibit these behaviors (i.e., genuine affective displays) may be viewed as ineffectual, boring, and uncaring.'" Once again, however, the LAPKA assessment penalizes this practice (i.e., a candidate is awarded zero points for mentioning "questions about feeling" in his/her response to Question 3, Scenario 1A).

Thus, as currently designed, it is hard to judge the LAPKA assessment as sensitive to differences across groups of teachers.

Appropriateness as a Method of Assessment

Teacher candidates and scorers largely agreed that LAPKA assesses knowledge relevant to elementary language arts instruction, but they expressed reservations about its appropriateness for assessing teachers' pedagogical competence. Their criticisms of LAPKA tended to fall into three categories: (1) LAPKA only measures teachers' ability to evaluate other teachers, (2) LAPKA measures what teachers say they do rather than what they actually do in the classroom, and (3) the true measure of good teaching is student learning, which LAPKA does not assess. Some of the teachers' comments are as follows:

> *This hasn't given any thought to see if I as a teacher in my own class can take techniques and adapt to my own class or if I can be successful in teaching it.*
>
> *It assesses my ability to intellectually evaluate a lesson but does not evaluate my actual teaching.*
>
> *Direct observation of teacher is better, or videotape the teacher.*
>
> *Success in teaching isn't measured on a piece of paper. The achievements and growth of students is success in teaching.*

The complaint by teachers that the LAPKA assessment does not evaluate "actual teaching" is a complaint often echoed by other teachers about other alternative assessments, especially those that are not performance-based. Perhaps if teachers had been asked to consider the appropriateness of LAPKA as one of several methods of assessing new teachers, their responses would have been different. Perhaps, too, if the teachers did not perceive the assessment as asking them to evaluate other teachers' instructional practices, but rather as asking them to demonstrate their knowledge of appropriate instructional practices in the areas of reading, writing, and literature analysis, the teachers would have considered LAPKA to be an appropriate method of assessment. It is the opinion of FWL staff, however, that it is unlikely that the teachers' perceptions of this method of assessment would change unless the content of the assessment were changed to better take advantage of the video stimulus.

Comparison with other assessments. In addition to being asked the question of appropriateness as a method of assessment, all of the teachers were asked the following: "How does this assessment format (i.e., answering questions after viewing videotaped lessons) compare with others with which you have been evaluated (e.g., multiple-choice for CBEST and NTE Speciality Area Tests, classroom observations during student teaching) in terms of its assessment ability?" While the teachers did not see LAPKA as appropriate as the sole measure of their teaching ability, they viewed it as a valuable supplement to direct observations and a significant improvement over multiple choice formats:

> *I like this format much better than multiple choice tests. I feel it is a better test of knowledge application. However, this type of test should not replace classroom observation.*
>
> *In comparison to the multiple choice section, you're able to explain reasons for answers. In comparison to classroom observations, you don't feel nervous and stress from being watched.*
>
> *[This is] better. The CBEST and NTE are so general as to be totally vague. They do not address the specific abilities and problems teachers face in actual classroom setting*

Only one candidate opposed the video format:

> *I feel this assessment rewards people who watch television and penalizes people who read.*

One teacher stressed that if we change the way teachers are assessed, then we also need to change the way teachers are taught:

> *If this is the way student teachers are tested then the curriculum needs to be adjusted to include more methodology and more practice scenarios.*

FWL staff concurs that, if this method of assessment were adopted by the state, the teacher preparation programs would do well to incorporate into their methodology classes the use of videotaped scenarios of teachers teaching.

Assessment Format

LAPKA's format is that of a written, constructed response assessment with a videotape stimulus. The assessment consists of four videotaped scenarios of teachers teaching, each of which is shown in segments. After each segment the teachers are asked to respond in writing to a series of questions about the material just viewed.

The format of the assessment is discussed by looking at the clarity of the following: (1) the materials sent to teachers in preparation for the assessment, (2) the assessment task materials (i.e., directions, literature, and questions), and (3) the scoring criteria and procedures.

Clarity of Teacher Preparation Materials

Prior to the assessment, the teachers received a two-page information sheet that included a brief explanation of the purpose of the assessment and logistical details (e.g., time, location), along with a two-page description of LAPKA supplied by the exercise developers. Most of the teachers indicated that they were satisfied with the materials that were sent to them.

When asked if there were any additional materials that would have been helpful to have in preparation for the assessment, the teachers made several suggestions. The most commonly made suggestion was to include some information about the scoring criteria. Another suggestion was to warn teachers to "be prepared to write a lot." Finally, one teacher thought the materials could be improved by providing more information "how the results will be used."

Clarity of Task Materials

The format for each of the four LAPKA scenarios included the following: "Instructions to the Viewers" given to the candidates at the beginning of each scenario to provide an "advance organizer" for the upcoming tasks, followed by a series of questions to be answered after viewing the scenario. In addition, for three of the four scenarios, teachers were provided the piece of literature which was the central focus of the videotaped lesson.

Although the teachers did not report any difficulty in reading or interpreting the literature selections for the three scenarios, the teachers did have some suggestions for improving the directions given at the beginning of each scenario and some of the assessment questions. These suggestions are presented below along with some suggestions made by FWL staff with regard to the assessment materials in general.

Suggestions for improving the task directions. Some of the teachers thought the directions preceding the scenarios could be improved by providing information that informed the viewer of the length and content of the upcoming video segment. Teachers reported being caught off guard several times by segments that were very brief, particularly some in Scenario 2, and as a result they felt unprepared to answer the questions. Commented one teacher,

> *Part F [in Scenario 2] was far too short! It would have been nice to be forewarned that it was approximately one minute in length.*

Teachers also reported being misled by the content of some the videos. In particular, in Scenario 3 some of the candidates reported that they expected the opening segment to include not only the teacher discussing her goals for the lesson but the actual lesson as well, and were surprised when the video was stopped and they had to answer the questions based only on the teacher's comments. These teachers' expectations were not unfounded as the prompt for Scenario 3 described the first video segment as containing "the teacher describing her objective and conducting the introduction to the lesson."

Based on FWL staff's observation of the confusion experienced by some teachers during the pilot test administrations and on our examination of the directions for the assessment tasks, we agree that the directions to the tasks could be improved by including the length of the segment to be viewed and an accurate description of the segment's content.

Suggestions for improving the questions. Not all of the questions in the scenarios were clear to the teachers and so there were several recommendations that the questions be reworded. One teacher explained,

> *Not always clear as to what you wanted responded to--wording not always clear.*

In particular, some of the questions from Scenarios 2 and 3 were identified as problematic. In Scenario 2, for example, Question 1, Part C asks the candidates to "identify the elements of an integrated approach to language arts which are present in this segment." Most of the candidates interpreted the word, "integrated," to refer to "integration across the curriculum" rather than the developers' intended meaning of "integration of the language arts" (i.e., reading, writing, speaking, listening). A typical comment was:

> *The integrated language arts video question was confusing because I have a different meaning for integrated language arts.*

Rephrasing the question to make explicit the concept of integrating the language arts would reduce the confusion.

In Scenario 3, some candidates experienced confusion with Question 1, Part A, which asked them to "identify and discuss two pieces of background information which are relevant to the lesson." The teachers were uncertain whether "background information" referred to their own background information, the teacher's background information for the lesson, or the students' background knowledge. A different word choice for "background information," or a rewording of the question, would probably take care of the problem.

Other suggestions for improving the materials. In addition to supporting the above teachers' recommendations, FWL staff suggest that consideration be given to the following recommendations. First, based on a review of the candidates' responses, the spaces for candidates' responses could be more clearly labelled. In some instances, questions were skipped over by candidates because the space for responding to them was not clearly marked (e.g., Scenario 1A, Question 3B on "One suggested improvement").

Second, we suggest that a review should be made of all the LAPKA questions so that only questions which are strictly dependent on viewing the videotape are included. In Scenario 2, for example, Question 2, Part B, asks the teachers, "Is 'publication' or 'sharing' an important part of the writing process? Why or why not?" This type of question can be answered without viewing the videotape and therefore does not take full advantage of the assessment method of video stimulus. Similarly, in the same scenario, another question asks the teachers to "comment on the strengths and/or weaknesses of this teacher's answer to the question on grading." A question of this type also does not take full advantage of the video stimulus because it could also be answered by the candidate by providing him/her with a script of the student's question and the teacher's response. Since the cost of videotapes, videotape equipment, etc. is much higher than paper-and-pencil assessments, the questions that are part of a video stimulus assessment should strive, as much as possible, to take full advantage of the video medium. That is, consideration should only be given to including questions that can not easily be answered through some other assessment format.

Third, although most of the teachers indicated that overall they had enough time to complete the assessment, several teachers stated that for Scenario 3 they would have liked more than the allotted time. Of more importance, however, is the fact that many of the teachers felt the entire assessment was too long. FWL staff suggests that consideration be given to the following teacher recommendations for reducing the time of the assessment: reduce the length of the videotape lessons; show fewer individual writing conferences; ask fewer questions; and eliminate any redundancies in questions.

Clarity of the Scoring Criteria and Procedures

Important concerns were identified by the scorers, our consultant on cultural diversity, and FWL staff regarding the LAPKA scoring system. These concerns, while overlapping to some degree, are presented as separate points.

Although the criteria and procedures for scoring the candidate's responses are quite straightforward--i.e., a candidate's responses are usually compared to a list of predetermined acceptable responses in the scoring key--the general consensus is that the scoring key is too narrow in scope and excludes many acceptable responses. In scoring the pilot tests, the scorers identified many responses that they believed exemplified effective language arts practices but were not credited because they were not listed on the scoring key. Some typical scorer remarks are as follows:

> *The scoring criteria are much too narrow in scope to allow for the variety of possibly valid answers.*

> *Answers would need to added that are just as appropriate as those now listed. The answers listed . . . reflect an interpretation of language arts pedagogy that is very limited.*

Because the scorers were advised by the trainers during the pilot test scoring session that their role was to score the candidate responses against the scoring key and not to add to or change it, the responses that fell outside the scoring key--but were perceived as valid by the scorers--were not discussed among the scorers and trainers during the scoring session. Thus, the scorers were given no opportunity to modify the scoring key.

Given the scorers' deep reservations about the narrowness of the scoring criteria, further development of this assessment might include a review of the pilot test responses and an examination of the responses that were not credited but that the scorers found to be acceptable practices. Through such an analysis, along with additional rounds of pilot testing and analyses, a more comprehensive list of acceptable responses might be developed. However, a potential problem could arise if through this process the list of acceptable responses becomes unmanageably long. A possible solution is to shorten the lists by combining many of the specific practices listed in the scoring key into a few general guidelines. This approach to scoring, however, requires a greater dependence on the scorers' professional knowledge and would require more extensive training to insure that the scorers understand and are applying the scoring guidelines in a consistent fashion.

The scorers also felt that too often the scoring key implied that there was a single "right answer," and ignored the influence of context on decision-making. One scorer expressed her concern as follows:

> *If there were true "right answers" to the myriad problems that teachers face everyday in their classroom, then school reform or even the reform of language arts teaching in elementary schools would be a simple matter of making sure everyone knew the "right answers." . . . There are ways of handling situations that are more or less appropriate for a given context, but no "right answers" that can remedy any situation, even the ones shown on the videotape.*

Another concern expressed by the scorers was that the assessment's analytic scoring focused their attention too much on details and not on the candidate's overall performance. Suggested one scorer:

> *I believe that some type of holistic scoring would be the answer. It would allow the candidate to be evaluated on his/her entire performance, rather than on tiny bits of it.*

A more holistic scoring system would perhaps address the scorers concern that the present scoring system does not allow scorers' sufficient latitude to apply their professional knowledge to make an overall judgment about a candidate's knowledge and skills. The scorers felt that they were required to apply the scoring criteria in too rigid a fashion, even when they had evidence from the candidate's response that the candidate lacked understanding of a particular concept or had misapplied the concept in the response. Allowing the scorers more latitude would enable them to look across a candidates' responses to find evidence of the candidate's understanding. Commented one scorer:

> *It frightens me that some candidates give an answer that fits within the parameters of acceptable answers and thereby get credit for it, but indicate elsewhere or even in the rest of their answer that they adhere to questionable language arts approaches.*

The scoring key was perceived by scorers as accepting "buzz words" as correct responses without examining the candidate's understanding of the concept underlying the word. One scorer remarked that the scoring system assessed teachers' "ability to use edu-speak."

The scorers also agreed with the teachers that some of the questions and directions were confusing, thus resulting in answers different from what the developers intended, and presenting a problem for scoring. One scorer remarked:

> *Another difficulty was scoring answers to questions that had obviously not been understood by the test-taker in the same way as they had been understood by the test-maker. [Also,] it was very difficult to score questions where the directions or layout of the questions apparently confused the candidate.*

Finally, other concerns about the fairness of the LAPKA scoring criteria and procedures with regard to different groups of teachers and teachers of diverse students were raised by our consultant on cultural diversity, Sharon Nelson-Barber. These concerns were already discussed in the Content section of this chapter, but to briefly restate her concern, the LAPKA scoring criteria and procedures seem to support one way to teach the language

BEST COPY AVAILABLE

arts to the exclusion of others. Nelson-Barber points out that this way is not necessarily appropriate for all students or teachers and that alternative approaches could also be appropriate. With the present scoring criteria and procedures, however, scorers who are aware of the many ways of teaching diverse student groups are unable to apply this knowledge.

Cost Analysis

Administration and Scoring Cost Estimate

The tasks of the Language Arts Pedagogical Knowledge Assessment are administered in a large group setting using video tape monitors to present teaching segments to which teacher candidates provide written responses. The tasks can be administered by one or more persons with little or no training in the specific content of the assessment using procedures common to standardized group test administrations.

Scoring requires the availability and training of raters knowledgeable in the content and criteria of the assessment. Scoring of the pilot test data, which included both training and actual scoring, involved two days for the six scorers to be trained and to score 42 teacher assessments. We estimate that, once trained, a rater could score approximately 20-30 assessments/day, providing the assessment consists of a set of tasks similar to the three scenarios pilot tested. Using $160/day for an scorer's cost would result in an estimate of $6.40/assessment (160/25) for scoring costs. We estimate that a two-day training session would be needed to train raters for this assessment. If we assume that 20 raters could be trained by one trainer in two days then the costs for 42 days of rater and trainer time at $160/day would equal $6,720. If it was assumed that each rater would participate in three days of scoring following this training then these training costs could be distributed across approximately 750 teacher candidates. The figure of 750 teacher candidates assumes that each of 20 scorers can rate 25 teacher candidates' responses each day and that each teacher's responses will be rated twice (i.e., 25 assessments/rater multiplied by three days multiplied by 20 raters equals 1,500). Dividing $6,720 by 1,500 results in an estimate of $5/teacher for training costs. This estimate could be increased or decreased as a function of the actual number of assessments that can be rated each day, the number of days an scorer rates after training, and the number of scorers trained at one time. But these figures are probably reasonable estimates of the costs for training and scoring assessments such as this. Combining $13 (i.e. $12.80 rounded off) for scoring and $5 for training costs results in an estimate of $18/teacher assessment for scoring.

As mentioned initially, the only special feature needed for administration is the video tape equipment. Adding $2 per assessment for these costs to the $30/assessment we have used for other similar administrations results in an estimate of $32/assessment for administration. A summary of cost estimates for administering and scoring an assessment like this is:

Training and Scoring:	$18/teacher
Administration:	$32/teacher
Total Scoring and Administration	$50/teacher

Development and Pilot Testing Costs

The costs for developing the tasks for this assessment were $115,528 and are broken out by major cost categories in Table 5.6, which also includes costs for pilot testing. These development costs are the expenses for the assessment developer to deliver the prototype activities to the CTC and SDE in the form they were used in pilot testing. $37,614 was spent for the pilot testing of these tasks with the 42 teachers. It is likely that future development and pilot testing that build on these and involve larger numbers of teachers would result in greater efficiency and lower costs. That is, it would not likely require the same level of development to obtain other video-taped scenarios and revised assessments; but since it is not possible to estimate the costs more precisely at this point, these data should provide a rough indication of the magnitude of effort for developing similar assessments.

Technical Quality

This section describes the technical issues related to the assessment.

Development

The Language Arts Pedagogical Knowledge Assessment was developed by Northwest Regional Educational Laboratory during the nine-month period from July, 1989 to April, 1990. The development team consisted of several staff members of Northwest Laboratory along with three outside consultants who are experts in the fields of assessment and/or language arts instruction.

TABLE 5.6

DEVELOPMENTAL AND PILOT TEST COSTS FOR THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT (LAPKA)

Cost Categories	Development	Pilot Testing
Staff-Salaries & Benefits	$36,495	$17,368
Consultants (Teachers, assessors, and other consultants)	14,400	6,352
Travel (Consultants and staff)	15,117	4,944
Other Direct Costs (Site rental, phone, duplication)	28,649	1,070
Total Direct Costs	$97,662	$29,734
Indirect Costs	17,866	7,880
Total Costs	$115,528	$37,614

213

In their initial meetings, the development team identified four scenario topics for videotaping: (1) A teacher-led, direct instructional activity, (2) an individual writing conference, (3) a small group, literature-based activity, and (4) a whole class discussion of a reading project. In addition, the development team decided to vary the grade level and student population across these scenarios, and that the teaching portrayed should model good practice.

In a subsequent session, the development team met with a group of practicing teachers who were identified as potential demonstration teachers to be videotaped for the scenarios. These teachers reviewed the scenario design plans and offered recommendations for developing the actual videotapes, but only one of the teachers agreed to be videotaped. Through additional recruitment efforts, the developers located teachers in San Jose and Oregon who were willing to be videotaped as demonstration teachers.

After the videotaping was completed in January, 1990, the development team reviewed the videotapes and edited them for assessment purposes. Contrary to the developers initial plans, however, the videotapes that were actually made, in some instances, did not reflect good teaching practices. In the final phase of the project, the development team prepared assessment questions and scoring criteria for each scenario.

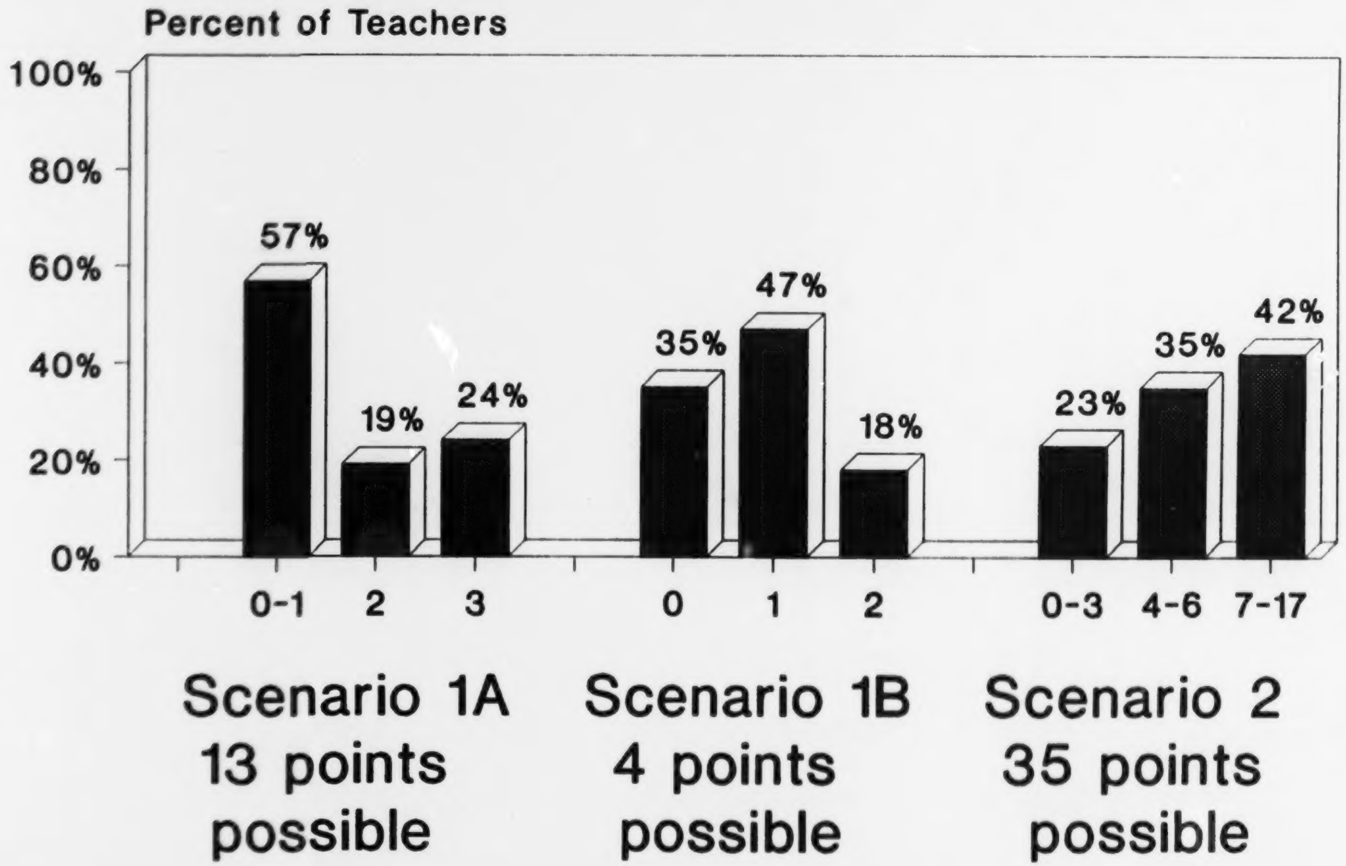
Reliability

The following analyses were performed on the pilot test data of 42 teachers. Interrater agreements were examined to assess the degree to which scorers were able to consistently judge candidates using the LAPKA scoring protocols for Scenarios 1A, 1B and 2 for which two raters rated each teacher candidate. Internal consistency estimates were generated to assess the degree to which scenario ratings would form a measure and the degree to which the different activities related to each other and might form an overall assessment.

Interrater agreements. Figure 5.1 contains a summary of the agreement between raters on Scenarios 1A, 1B, and 2, for which there were two ratings. The agreements were modest on Scenario 1A where 57% of the candidates were assigned equivalent scores within the 13-point range for that scenario. Scenario 1B has a range of only four points. Thirty-five percent (35%) of the candidates were assigned the same ratings and 47 percent of the ratings differed by one point on Scenario 1B. Scenario 2 has a score range from 0-35 points. Only 23 percent of the candidates were assigned scores within three points. Forty-two percent differed by seven or more points. These results support the need for additional

FIGURE 5.1

% Agreement of Raters for the Language Arts Pedagogical Knowledge Assessment



revisions and development in the scoring system. The scorers and teacher candidates provided helpful feedback on concerns with the current scoring system and FWL has taken these and our evaluations into account in making suggestions for improvements and next steps. Many of these have been described earlier and will be highlighted in the summary and conclusions section which follows.

Interrater correlations. The correlations between raters are summarized below:

	Rater Pair			Averaged
Scenario	1	2	3	Pair Ratings
1A	.22	.91	.16	.56
1B	-.25	--	.00	NA
2	.46	.42	.46	.45

Interrater agreements were low and the ratings on 1B precluded computing correlations for two rater pairs, e.g. in Rater Pair 2 one rater assigned all teachers the same rating for the five teachers s/he rated. Given the small N's and needs for further development in the scoring system, these estimates should be interpreted as only preliminary and lower bound estimates of the potential for obtaining consistent ratings with this assessment prototype.

Internal consistency of the scenarios. Coefficient Alpha reliability estimates were calculated for the scenarios and are presented below:

Scenario	Reliability
1A	.45
1B	-1.09
2	.25
3	.53

These estimates are also quite low and imply that the ratings conducted in this pilot test achieved little internal consistency. This is probably attributable to several factors. One of which is the small sample size associated with the pilot test. But equally important are the issues related to the factors described earlier. For example in Scenario 1B the "negative" reliability derives from the fact that two items negatively correlated with the third on a three item task. The interrater and internal consistency data provide empirical evidence of the need for further development, refinement, and research before this assessment would be ready for large scale administration or field testing.

Intercorrelations among tasks. The data above on reliabilities provide evidence to predict that there will not be a very high correlation among the scenarios given the low reliabilities of the current measures. For informational purposes we have listed correlations among tasks below:

	IA	2	3
IA	--		
2	-.20	--	
3	-.12	.41	--

	IB	2	3
IB	--		
2	.23	--	
3	-.12	.41	--

The .41 correlation between Scenarios 2 and 3 is the only statistically significant relationship and this is partially accounted for by the fact that it is based on the 39 teachers, whereas correlations involving Scenarios 1A and 1B were based on only half the teachers.

Validity of Agreement Through Group Comparisons

Appendix D contains means, standard deviations, and related statistics for teacher candidates by several variables (i.e., gender, number of courses, grade level taught, teaching location: urban-innercity-suburban, and minority-nonminority). There were no notable trends in these data. This is again a combination of the fairly low reliability of the measures and small numbers in some groups. For example, the similarity in minority and nonminority teacher performance in this pilot test is simply an indication of the assessment's weak ability to reliably assess teacher candidates.

Content Validity

Evidence of the content validity of the assessment comes from three sources. The first is the role that the developers and the experts in the fields of assessment and language arts instruction, had in designing the assessment. The second is the analyses of the match of the assessment to the model curriculum guide and California Beginning Teacher Standards that compare the assessment's content with that recommended in the official documents. The third is in the type of concerns raised by the beginning teachers who participated in the pilot test. These have been described earlier and implications for further development are addressed in the following section.

In summary, the current scoring criteria and training with this pilot test sample did not produce sufficient data on which to judge accurately the technical quality potential of this prototype assessment. Perhaps, with revisions in, and further development of, the scoring criteria, future pilot tests could yield data that will support the technical quality and potential of the assessment.

Conclusions and Recommendations

Language arts instruction, the focus of this assessment, is central to elementary school teaching. As an instrument for measuring teachers' pedagogical content knowledge in this area, LAPKA has some strengths as well as some serious weaknesses, particularly in the area of scoring. In light of these strengths and weaknesses, FWL offers a number of recommendations for modifying this instrument.

Administration of Assessment

LAPKA is relatively easy to administer and can be efficiently given to large groups of teachers. Other than the need for videotape players and monitors, the requirements for administration are similar to other large-scale, paper-and-pencil tests. LAPKA requires no specialized knowledge from test administrators.

LAPKA consists of four different scenarios, each of which requires approximately an hour and a quarter to complete. For the pilot test, candidates took three of four scenarios during a four-and-a-half hour administration session. Based upon the candidates feedback, we recommend the following:

- Limit the assessment to no more than three hours. Many candidates indicated that fatigue began to interfere with their performance.

Orientation materials for this pilot test consisted of a two-page description of the assessment. Based on teachers' feedback about these materials, we recommend the following:

- Provide candidates with more detailed orientation materials, including a description of a sample scenario, sample questions, and examples of the scoring criteria. In addition, it might be appropriate (and more efficient) to provide candidates with the literature selections ahead of time if advance knowledge of the text would not interfere with the goals of the assessment.

Based on projected cost-per-candidate, LAPKA is relatively inexpensive to administer. However, each time that LAPKA is administered, a new set of videotapes and scoring criteria will need to be developed, which will add to the overall cost-per-candidate. Therefore, we recommend the following:

- Limit the number of administration dates to minimize ongoing development costs.

Ensuring that all candidates are given adequate opportunities to demonstrate what they know and can do is one critical feature of an assessment; employing scorers who understand and can recognize diverse ways of "good teaching" is another critical aspect of creating a fair and effective assessment. Thus, we recommend the following:

- Recruit scorers who are recognized for their teaching excellence in language arts and who are knowledgeable of the California frameworks and curriculum guides in English-language arts.

- Recruit scorers who have knowledge of and experience in multiple cultural settings.

Based on the scorers' feedback, we recommend the following changes to the scorer training program:

- Lengthen the time for training from the present six hours to a two day session;

- Increase the amount of practice materials for scorers to use during training, with examples of poor and acceptable responses, and a range of responses that reflect diverse effective practices;

- Allow teachers to view the videotapes from beginning to end before they actually begin to score the practice materials.

Assessment Content

The conclusions presented below are based on feedback collected from the pilot test teachers, scorers, and a consultant to the project on issues of diversity and equity, along with the observations of FWL staff.

- Overall, LAPKA's congruence with the California English-Language Arts Model Curriculum Guide for Kindergarten through Grade Eight is good. While LAPKA does not address all of the guidelines, it portrays a variety of reading, writing, and oral language activities, all of which are consistent with the curriculum guide.

- Coverage by LAPKA of the California Standards for Beginning Teachers is limited. A small number of standards are addressed, and these standards are addressed only indirectly because LAPKA does not provide direct evidence of a teacher's ability to perform in the classroom. Adding questions that specifically address the standards could improve the information for the assessment to better address the standards.

- LAPKA appeared to be moderately difficult for beginning teachers. The candidates correctly answered from 57% to 87% of the questions in each scenario. The candidates had the most difficulty with Scenario 2, which also had the greatest number of confusing questions. Eliminating or rewording these questions would possibly improve candidates' scores. The range in candidate scores suggests that this exercise might be effective in discriminating weaker from stronger candidates in the areas assessed.

BEST COPY AVAILABLE

- Teachers and scorers thought that the assessment was fair to teachers of different grade levels. An analysis of the performance data indicates no significant differences between teachers who taught in grades K-3 and 4-8.

- The LAPKA scenarios portray students of diverse cultures and various ability levels. Although most of the teachers perceived this assessment to be appropriate for teachers of diverse student groups, some expressed concern about the lack of portrayal in the scenarios of bilingual classes, special education students, and limited English speakers. In addition, as pointed our by our consultant on cultural diversity, LAPKA tends to reward only certain ways of teaching and, as a result, has the potential of discriminating against certain teaching techniques deemed effective in some minority communities. Thus, it is the conclusion of FWL staff that, as presently constructed, LAPKA may not be effective for assessing teachers across contexts and with diverse students.

- While most teachers though LAPKA was fair to different groups of teachers, this view was not shared by some of the teachers, the scorers, and the consultant on cultural diversity. One concern was the lack of teacher diversity shown on the videotapes (e.g., few minorities, no males). Another concern was that the LAPKA scoring process fails to recognize a range of culturally diverse ways of teaching.

- The performance data indicated no significant differences in candidate performance based on gender, ethnicity, grade level taught, California or non-California teacher training program, or the number of reading methods courses taken.

- The majority of teachers and scorers felt that LAPKA is not an appropriate way to assess teachers. While they offered a variety of reasons for this view, the most frequent objection cited was that LAPKA does not assess teachers' ability to teach, instead it assesses their ability to evaluate other teachers.

Based upon the teachers' and scorers' comments, as well as our review of the content of the assessment, we recommend the following changes to the individual scenarios:

222

- Reconceptualize the writing activity in Scenario 2. The teachers and scorers felt that the individual writing conferences in this scenario were quite unrealistic. In a class of twenty-five to thirty students, teachers rarely get the opportunity to sit down with individual students and hold extended conferences about a particular piece of writing. However, teachers do respond in person to students individually--but usually in small group settings or at brief, desk-side interviews. In small, peer-editing groups, for example, students (and the teacher, if present) respond to the possibilities and problems in a student's draft. In fact, some teachers argue that this small-group context is not only a more efficient way to respond to individual students, but a more effective one as well.

- Begin Scenarios 1A and 1B with the videotape teacher presenting her goals for instruction--as was done with Scenario 3. Scenarios 1A and 1B differ from Scenarios 3 and 4 in several significant ways. Scenarios 3 and 4 begin with the teacher explaining her goals for instruction to the viewer, while Scenarios 1A and 1B begin with the actual lesson. Given that many teaching practices can be appropriate or inappropriate depending upon the teacher's goals, contextual factors, etc., it seems particularly important for the viewer to be aware of some of these features--in particular, the teacher's goals for the lesson or unit. It is much more difficult to evaluate a teacher's instruction without knowing that teacher's goals for the lesson or unit. In addition, for all of the scenarios, it might also be productive to provide the viewer with some information about the students, the course, etc.

Assessment Format

The LAPKA format has a number of strengths: (a) The video stimulus can present the contexts and complexities of teaching in ways that paper-and-pencil formats cannot; (b) The open-ended question format, in which candidates construct brief written responses, gives candidates the freedom to focus on particular aspects of the videotaped teacher's performance as well as the responsibility for appropriately framing their responses, and (c) Requiring candidates to give rationales for many of their responses allows a candidate's reasoning to be taken into account in scoring his or her responses.

Despite these strengths, however, the format of LAPKA could be improved in numerous ways. First, based upon the perceptions of the teachers, scorers, and FWL staff, we believe that consideration should be given to the following recommendations:

- Prior to showing a video segment, provide information to the viewers about the length and content of each segment.

- Improve the clarity of the scenario questions which caused confusion for the pilot test participants.

- More clearly label the spaces provided for candidates' responses.

- Review all LAPKA questions so that only those questions which are strictly dependent on viewing the videotape are included.

- Consider including only those questions which take full advantage of the videotape stimulus and cannot easily be answered through some other assessment format.

- Reduce the length of the assessment by reducing the length of the videotape lessons, showing fewer writing conferences, asking fewer questions, and eliminating any redundant questions.

Finally, to improve LAPKA's format, consideration needs to be given to substantially changing LAPKA's scoring process. In response to the reservations expressed by the scorers, the consultant on cultural diversity, and FWL staff about LAPKA's scoring criteria and procedures, we recommend the following:

- Broaden the range of responses considered "acceptable" in the scoring key, especially taking into account culturally diverse conceptions of teaching. This might be achieved by reviewing the pilot test responses and adding additional rounds of pilot tests, particularly with minority candidates, to build a more comprehensive picture of the range of acceptable responses;

- Take a candidate's teaching framework into account as explicitly as possible in scoring a candidate's responses. This might be achieved by having candidates write a brief statement of philosophy or goal statement for each scenario (e.g., views on the teaching of writing), and by having candidates provide rationale statements with their responses (done to some extent in the present version of LAPKA).

- Move from the present analytical scoring system to a more holistic one that would allow scorers more flexibility to look across a candidate's entire performance and that would enable scorers to consider a candidate's responses in light of the candidate's rationales and philosophy. This shift would require a number of changes in the scoring system (e.g., document their reasons for assigning a particular rating).

In addition to the major changes described above, we believe that the scoring format would be further improved by the following these recommendations:

- Provide an explicit rationale for why some responses are acceptable while others are not;

- Provide a rationale for why some responses are awarded two points, other responses one point;

- Develop a plan for combining scores across scenarios to determine a total LAPKA score;

- Provide a tentative estimate describing the relationship between candidate scores and levels of proficiency in teaching elementary language arts (e.g., weak, adequate, exemplary).

Summary

LAPKA, an assessment of pedagogical content knowledge in the elementary language arts, is easy and relatively inexpensive to administer, captures some of the complexities of teaching through its video format, and assesses candidates' knowledge on a wide range of language arts practices. The present scoring system, however, requires substantial revisions. In particular, the scoring system needs to be more sensitive to the context-dependent nature of instructional decision-making and to conceptions of teaching different from those portrayed in the videotapes.

CHAPTER 6:

STRUCTURED SIMULATION TASKS FOR SECONDARY ENGLISH TEACHERS

The Structured Simulation Tasks for Secondary English Teachers, developed by the RAND corporation, are a set of structured simulation problems to which a teacher responds in writing. The assessment resembles the Structured Simulation Tasks for Secondary Life/General Science teachers, also developed by RAND (and described in Chapter 3), in that it does not constitute a complete assessment. Instead, development of the assessment focused on construction of prototypic tasks, some or all of which, combined with other forms of assessment, could be used to evaluate English teacher candidates.

Because the Structured Simulation Tasks were developed as discrete assessment exercises and not as a series of exercises that comprise a single assessment, this evaluation of the tasks is not an evaluation of a single assessment. (To simplify references to these prototypic tasks, however, they are sometimes referred to collectively as "the assessment.") This chapter describes and evaluates the strengths and weaknesses in the administration, content and scoring of the individual assessment exercises rather than the assessment as a whole. Perhaps because the Structured Simulation Tasks for Secondary English Teachers are very different from the Structured Simulation Tasks for Secondary Life/General Science Teachers, the strengths and weaknesses of the former are generally different than those described in Chapter 3 for the latter.

The chapter is organized into the following sections: The Structured Simulation Tasks, Administration of Assessment, Scoring, Task Content, Assessment Format, Cost Analysis and Technical Quality. A synopsis of findings is outlined in the Summary, Recommendations, and Conclusion sections at the end.

The Structured Simulation Tasks

The Structured Simulation Tasks for Secondary English Teachers were designed to elicit demonstrations of teacher knowledge specific to the content area of English/Language Arts. Unlike traditional multiple choice tests, responses generated from simulation tasks are generally open-ended, allowing candidates to bring to bear relevant insights from a range of knowledge domains. Five paper and pencil simulation tasks were developed, scored and evaluated for the pilot test:

(1) **Responding to Typical Problem Situations** (90 minutes) This task assesses teachers' ability to suggest workable solutions to problems that commonly confront English/Language Arts teachers. Teachers are presented with six vignettes or scenarios that outline hypothetical problems. Problem situations include taking over a class mid-semester, dealing with parent complaints, handling plagiarism and cheating, and facilitating ESL instruction. After reading the problem situation, teachers write one or two paragraphs, or a list of bulleted ideas, regarding one or more alternatives to addressing the situation presented.

Two forms of **Responding to Typical Problem Situations** were developed and administered. Both forms share one scenario in common; the other five scenarios parallel each other with similar types of problems presented, although specific details about students (i.e., grade-level) or parents involved differ.

(2) **Designing a Lesson Sequence** (120 minutes) This task assesses teachers' ability to design a five-day lesson plan. Teachers are given two literary essays selected from different genres, unit objectives, and contextual information. Teachers then create a five-day lesson plan that instructs students in the themes of the two works, and prepares them to complete a first draft writing assignment. For each lesson, teachers must outline objectives, activities, and a rationale explaining how activities achieve their goals.

(3) **Responding to Student Writing** (60 minutes) This two-part task assesses a teacher's ability to 1) respond to first draft writing in a manner that encourages and guides students through the writing process, and 2) identify strengths and weaknesses of written teacher responses to students' first-draft writing. In part I, teachers are given two student papers to comment on. Teachers comment on the papers as if they are first drafts. In part II, teachers critique comments a hypothetical teacher has made on three other student essays written on the same topic as in Part I. In both parts, teachers are provided background information about the assignment that generated the papers and the students who wrote them.

(4) **Stages of the Writing Process** (60 minutes) This task is designed to assess a teacher's ability to 1) identify strengths and weaknesses in a sequence of classroom activities whose purpose is to promote student writing, and 2)

develop and revise activities that would contribute to improving student writing skills. The task has two parts. Teachers are first presented with an assignment followed by a series of eight activities designed to take students through the brainstorming, drafting, and revising of an essay. Teachers critique the assignment and activities for strengths and weaknesses. In part II, teachers suggest ways they would design the assignment and/or activities differently.

(5) **Developing Oral Presentation Skills** (80 minutes) This task assesses teachers' ability to identify strengths and weaknesses of a hypothetical teacher's actions and comments during an oral presentation activity. In the first of this two-part task, teachers critique the strengths and weaknesses of an assignment that focuses on small group oral presentations. In the second part, teachers are presented with two realistic scripts, each "excerpted" from class sessions in which student groups present oral assignments. Candidates critique strengths and weaknesses of teacher statements and actions portrayed in the two scripts.

Administration of Assessment Tasks

Following an overview of the administration of the assessment tasks, this section contains information on the following: logistics (e.g., identifying the teacher sample, administering the tasks), the teacher sample, assessors, security and teacher perceptions of the assessment administration.

Overview

The Structured Simulation Tasks for Secondary English Teachers were administered to 56 teachers during the spring of 1991 at five sites throughout California, including the San Francisco Bay Area, the greater Sacramento and Los Angeles areas, and the Imperial Valley. Since completing all tasks could take over eight hours, the tasks were divided into two groups of three. Twenty-eight teachers took **Typical Problem Situations (Form A)**, **Responding to Student Writing**, and **Designing a Lesson Sequence**. For convenience, we will refer to these teachers and tasks as Group I. The other 28 teachers completed the remaining three tasks--**Typical Problem Situations (Form B)**, **Stages of the Writing Process**, and **Oral Presentation Skills**--which we will designate as Group II. For each group, two tasks were administered in the morning, and one task was administered after lunch.

Logistics

Conducting the pilot study included the following activities: identifying teacher samples, sending orientation materials, administering the tasks, and collecting evaluation feedback from the participating teachers.

Identifying teacher samples. Most of the 56 teachers who piloted the structured simulation tasks were identified based on their participation in the California New Teacher Project (CNTP). Not all of those asked to participate did, however, mostly for logistical reasons (e.g., the scheduled administration date conflicted with a previously set engagement). Thus, a small proportion of teachers were recruited who were not CNTP participants. These teachers were recruited from school districts that neighbored CNTP projects.

Sending orientation materials. The assessment developer provided the orientation material for the teachers, which consisted of brief descriptions of the six possible tasks they would be asked to complete. In addition, teachers received a letter briefly describing the California New Teacher Project and its assessment component, as well as directions to the assessment site. Teachers were paid $150 for participating in the assessment and completing an evaluation feedback form.

Administering the tasks. The assessment tasks were designed to be administered to large groups by a test administrator who distributed and collected materials, announced the start and end of each task, and monitored the teachers to prevent cheating. No special training or background in secondary English was needed, as the task instructions were designed to be self-evident.

The only requirement which differed from those of traditional group-administered tests was that of sufficient surface area (e.g., individual desks or a number of tables) to spread out a number of materials. Facilities which fit this requirement proved to be easy to locate, and included university classrooms, classrooms at a district professional development center, and a classroom at a church center.

Each administration began with a 10-15 minute overview of the research design underlying the California New Teacher Project. Teachers were then directed to open the first of three manila envelopes placed before them, each of which contained the test materials for a single task. After completion of the first task, teachers were given the option of a five- to fifteen-minute break, and usually opted for five minutes. After the second task, teachers took a lunch break. The third task was administered after lunch.

Collecting evaluation feedback. Upon completion of the third task, teachers were asked to fill out a feedback form in which they were asked to give us their perceptions of the administration of the tasks, as well as the tasks themselves.

The Teacher Sample

The background characteristics and teaching contexts of the teacher sample for this study are summarized in Table 6.1 and discussed in the following two sections.

Background characteristics and preparation. Eighty-two percent of Group I teachers and 62% of Group II teachers were female. An overwhelming proportion in both groups--86%--were Anglo. Just over half the teachers in each group held undergraduate majors in English, with about 75% in each group having taken at least one English methods course. About 85% of the teachers had taught for up to two years full-time. Seventy-five percent of the teachers held a single subject credential in English, while 15% of Group I teachers and 20% of Group II teachers held emergency credentials.

Teaching contexts. In Group I, 64% of the teachers taught grades 6-8, compared to 18% in Group II. Conversely, 36% of Group I teachers taught grades 9-12, compared with 82% in Group II. Whereas 75% of Group I teachers taught in urban or inner-city contexts, 75% of Group II teachers taught in predominately rural or suburban settings. In spite of the different settings in which teachers work, the distribution of languages spoken by teachers' students in each group was similar and ranged from two to more than nine.

Security

It is the position of the test developer that once the tasks are given, assessment security is compromised and new forms of the tasks must be developed. The "shell" system used to develop these tasks (see **Development**) potentially enables parallel tasks and activities to be designed for different administrations. Such a system can minimize the potential for candidates to memorize acceptable answers on some tasks, such as **Responding to Typical Problem Situations** and **Stages of the Writing Process.**

TABLE 6.1

PILOT TEST PARTICIPANTS

STRUCTURED SIMULATION TASKS FOR SECONDARY ENGLISH TEACHERS

(Number of Teachers = 56)

Descriptive Characteristics of Participants	Distribution of Participants	
	Group I (N=28)	Group II (N=28)
Gender		
Male	4	9
Female	24	19
Ethnicity		
Asian/Pacific Islander	2	0
Black	0	1
Hispanic	2	3
White	24	24
Undergraduate Major		
English	21	20
Humanities/Liberal Arts	6	5
Science	0	3
Other	1	0
Grade Level Taught		
6th - 8th Grade	18	5
9th - 12th Grade	10	23
School Setting		
Rural	2	8
Suburban	4	13
Urban	8	7
Inner-city	14	0

Assessors and Their Training

Members of FWL staff administered the assessment tasks. No training was provided, other than instructions about times for the tasks and suggested breaks. Before the first administration, staff members designed the assessment schedule, including time to provide an overview of the tasks and the goals in developing them, and time afterwards to complete an evaluation form. No need for further training was required, as members of FWL were experienced in administering similar kinds of assessments.

Teacher and Assessor Impressions of Administration Logistics

Overall, the teachers responded favorably when asked their impressions of the arrangements for administration, including scheduling, room arrangements, and distance to travel to the assessment site. All but one of the 28 teachers in Group I found the arrangements to be reasonable, and in Group II, 25 of the 28 teachers were satisfied with the arrangements. Two of the three teachers who were not satisfied had criticism for the room arrangements, saying that the one assigned was too small and that some teachers had to be moved to another room for lack of chairs.

Scoring

Scoring of the tasks revealed the greatest challenges in developing structured simulations as viable means of assessment. The following section focuses on scoring logistics, scorers and their training, and the scoring process. Results of the scoring are discussed in the section **Performance on Structured Simulation Tasks**, in the **Content** section.

Logistics

Scoring took place at the RAND Corporation in Santa Monica over a two-week period in June 1991. In addition to two RAND developers, eight scorers participated during the course of the two weeks. During the first week, two teams, comprised of three raters each, scored four tasks; during the second week, another two teams, also comprised of three raters each, scored the remaining two tasks. After all tasks had been scored, a fifth team, comprised of four second-week scorers, re-scored two of the tasks as a check on scorer reliability.

Scorers and Their Training

As noted above, there were ten scorers altogether: two were RAND developers, and eight were recruited by FWL based on referrals from the California Department of Education. In all, three scorers were female, seven were male. One of the women was Black, one of the men was Hispanic; otherwise, all scorers were Anglo. Scoring teams were created based on the career backgrounds of scorers. On each scoring team sat a RAND developer, a district administrator and/or classroom teacher, and, in four of the five scoring teams, a university professor.

Each of the two RAND developers supervised and participated in the scoring of all tasks. Though one had university teaching experience (in experimental methods), neither had a background in secondary classroom teaching or school administration. Of the remaining eight scorers, one had served as a member of the test development committee, though not as a RAND employee. This scorer had extensive experience both in high school English classrooms and as a district administrator. Another scorer was a district administrator with extensive experience at the elementary level (mostly grades 4-6), both in and out of the classroom. Three of the scorers were high school English teachers, and two of these were also heads of their English departments. Finally, three scorers were university professors, either of English, Composition, or Teacher Education. Two of the university professors had experience teaching secondary level English.

None of the scorers were trained before actual scoring began. At the start of the first day of each week, RAND developers and FWL evaluators explained the process and goals of the structured simulation tasks. Then, RAND developers chose a task to score. There was no training as to the meaning of scoring criteria or in how to apply the criteria. There was also no training in a procedure by which to calibrate scoring before actual responses were rated. According to the developers, this was in large part due to the fact that there were so few responses per task--no more than 28. RAND developers note that establishing reliability among raters for similar tests, such as the bar exam, usually takes fifty samples, which was clearly not possible at this stage of development.

Scoring Process

As was the case with the science structured simulation tasks, analytic scoring guides were built into the English structured simulation tasks; in fact, in some cases, scoring guides were developed before the tasks. In analytic scoring, pre-determined criteria are compared to each teacher's response. Each criterion specifies whether one or more points will be awarded or deducted. Elements of a teacher's response are then matched with scoring guide criteria, and points awarded or deducted accordingly (see Figure 1 for an example of an analytic scoring guide).

Unlike the scoring of the science structured simulation tasks, however, the scoring of the English structured simulation tasks did not strictly adhere to the analytic scoring method. Instead, during the actual scoring of the tasks, the analytic method was sometimes abandoned or greatly modified in favor of the holistic method. An alternative to analytic scoring, holistic scoring is based on the quality of a response taken as a whole. There are two types of holistic scoring: normed and criterion-referenced, or focused, holistic scoring judges responses in relation to pre-selected responses which conform to specific criteria and represent rating points on a scale. Criterion-referenced scoring is preferred for assessments where the comparability of standards across assessment administrations is important.

When the scorers opted for the holistic scoring method to evaluate some tasks or parts of tasks, they chose normed holistic scoring and used the following procedure: The first response was judged for overall quality and then placed in a pile. The next response was judged relative to the previous one, then placed in the same pile, a higher scoring pile, or a lower scoring one. In this way, a rough scoring scale was developed as the responses are read. When all responses were judged, there were from two to ten piles, each representing a different grade of quality. In making the judgements, scorers relied partly on the criteria outlined in the original scoring guides; however, decisions were largely based on two factors. The first was professional experience, which sometimes encompassed criteria beyond those in the scoring guides. The second factor considered was the relative quality of a particular response compared to those that have immediately preceded it.

It wasn't until actual responses were on the table that the effectiveness of the pre-determined criteria could be evaluated. As it turned out, the analytic scoring guides were used to differing degrees across all tasks and activities within tasks. In a few cases, scoring guides were used as constructed. In other cases, they were altered, added to, or abandoned in favor of different criteria altogether. In most cases, the analytic scoring guides became guidelines by which to assess responses holistically.

Members of scoring teams preferred holistic scoring when presented the option, especially for more open-ended simulations such as **Responding to Student Writing** and **Designing a Lesson Sequence**. Holistic assessment was intuitively more satisfying to raters because it allowed them to exercise professional judgments about overall quality. Raters resented giving one point for both a trivial as well as a sophisticated perception. Although this problem was alleviated in a few cases by weighing some scoring criteria more heavily, a respondent who listed a number of trivial points could still earn a score equal to a teacher who listed fewer but more important insights. Important differences in scoring outcomes resulted from relying on one or another method. These differences are discussed in the section, **Performance on Structured Simulation Tasks**, under the sub-section, "Analytic versus holistic scoring."

In the following sections, the scoring method used for each structured simulation task is discussed. For discussion of results, see **Performance on Structured Simulation Tasks**.

Responding to Typical Problem Situations. There were eleven scenarios piloted, six for each of the two groups of teachers, one of which was responded to by both groups.

Seven of the scenarios were scored analytically, using the scoring guides provided. Teachers earned one point for matching a criterion on the scoring guide, or had one point deducted for inappropriate suggestions. Provisions existed for raters to award or deduct points for responses not listed on the scoring guide. Overall scores on the analytically scored scenarios ranged from -2 to +4, though the range for each scenario varied.

Five scenarios were scored holistically, based to varying degrees on the scoring guides provided (one of the holistically scored scenarios was re-scored by a different scoring team using the analytic method). Scales of holistic scores ranged from 1 (low) to 3, 4, or 5 (high).

Designing A Lesson Sequence. In this task, teachers were asked to develop a five-day lesson sequence, based on materials and objectives provided in the task materials. The original scoring guide directs scorers to evaluate each day's lesson plan in terms of a prescribed list of criteria. However, scorers felt that focusing on the lesson plans day by day would fail to yield a meaningful assessment of a teacher's overall approach to a unit. For example, one of the important items on the day-by-day scoring criteria had to do with how well teachers fulfilled the weekly objectives. Both the initial scoring team and one that re-scored the exercise felt that a teacher's ability to fulfill unit objectives is best looked at across the span of the entire unit, not within each lesson plan. As a result, both scoring teams shifted to a holistic assessment of responses, assigning a single score for the unit, instead of one score for each of the five days. Both scoring teams used the scoring criteria provided as a guide, as well as adding a few of their own. The original scoring team's range was from 1-5 (5=high); the re-score team's range was 1-7 (7=high).

Responding to Student Writing. There were two parts to this task. Part I presented teachers with student papers (with some contextual information) upon which to comment. This part of the task was scored holistically, based in part on scoring guide criteria (Essay #1 range = 1-6, 6=high; Essay #2 range = 1-5, 5=high). Part II presented teachers with papers commented upon by a hypothetical teacher and asked teachers to identify strengths and weaknesses in the comments. This part was scored analytically, with modifications made to the original scoring rubric both in terms of criteria to be used and the weighing of those criteria: some criterion were worth 1 point, while more important ones were worth up to 2 points. Teachers lost points for misidentifying a strength or weakness. The range of scores on Essays 3, 4, and 5 was -1 to 6, -2 to 7, and 2 to 9, respectively.

Stages of the Writing Process. There were two parts to this task. In Part I, teachers identified strengths and weaknesses of a writing assignment and eight activities leading to a completed essay. This section yielded nine scores altogether. In Part II, teachers were asked how they might have structured the assignment or tasks differently. This section yielded one score.

Three methods of scoring were used across the ten sub-tasks. One sub-task was scored analytically, using the scoring criteria provided. For some other sub-tasks, holistic scoring was used, based to varying degrees on the original scoring guides. Finally, some sub-tasks were scored using a combination analytic and holistic scoring method. Raters would read a response and mark points, then announce their point totals. Based primarily on points and some qualitative considerations, responses were sorted into categories ranging from poor to excellent. Where holistic scoring was used, category 1 was always lowest, with the upper category being 2, 3, 4, and most often 5.

Developing Oral Presentation Skills. This task had two parts, three sub-tasks altogether. In Part I, teachers assessed the strengths and weaknesses of an assignment designed to foster students' oral presentation skills. In Part II, teachers were presented with transcripts of two class sessions where oral presentations are made by students. Teachers were to evaluate strengths and weaknesses of a hypothetical teacher's comments and behaviors in each of the two transcripts. Both Part I and Part II were rated analytically using the scoring criteria provided, with some modifications to eliminate overlap of criteria and to account for answers not originally anticipated. Teachers earned a point for matching strengths and weaknesses in scoring criteria, or had a point deducted for calling a strength a weakness or vice versa. Scores on Part I ranged from 0 to 9; Part II, segment 1, from 2 to 15; Part II, segment 2, from 3 to 24.

Structured Simulation Tasks' Content

In the following pages, the content of the simulation tasks is evaluated along these dimensions:

- Coverage of the 1987 California English/Language Arts framework;
- Extent of coverage of the California Standards for Beginning Teachers;
- Job-relatedness of the tasks;
- Appropriateness of tasks for beginning teachers;
- Performance on structured simulation tasks;
- Appropriateness across contexts (e.g., grade levels and students taught);
- Fairness across different groups of teachers (e.g., gender, ethnicity); and
- Appropriateness as a method of assessment.

Coverage of California English/Language Arts Framework

The California State Department of Education periodically produces subject-specific documents, curriculum guides, and frameworks, which serve as public statements describing the curricula that content and pedagogy experts believe are most appropriate for California's youth. The most recent document pertaining to English and language arts instruction is the *English/Language Arts Framework* (California State Department of Education, 1987). This framework suggests overall approaches teachers ought to use in designing and implementing curriculum. Rather than specify specific works to study or appropriate times in students' development to introduce certain skills or knowledge, the framework provides an educational philosophy and some specific guidelines for teachers to follow in designing and implementing curricula.

Key elements of the *English/Language Arts Framework* that are pertinent to secondary level English teachers can be derived from pages three and four of the framework. In the following pages, each element (italicized) is discussed in relation to the coverage it received in each task (see Table 6.2). Since "coverage" refers to the extent to

TABLE 6.2

COVERAGE OF ENGLISH/LANGUAGE ARTS FRAMEWORK

Element of the Framework	Responding to Typical Problems A and B	Responding to Student Writing	Designing a Lesson Sequence	Stages of the Writing Process	Developing Oral Presentation Skills
Structuring a literature-based curriculum	L	L	P	L	L
Implementing a meaning-centered curriculum that promotes critical thinking and integrates skills of reading, writing, listening, and speaking	P	P	F	L	L
Attending to various "stages" of the writing process	P	F	L	F	N
Sensitizing students to values in literature that "reflect real dilemmas faced by all human beings..."	P	L	P	N	N
Designing instructional programs that provide opportunities for critical thinking among all students, regardless of prior preparation or language ability	P	P	P	L	P
Developing oral language skills among all students	L	N	L	L	F
Contributing to a school environment where students are encouraged to read widely, write frequently, and listen and speak effectively	P	N	L	N	N
Modeling effective use of all language arts skills	N	N	N	N	N
Devising an assessment program that encompasses the full range of language arts goals that incorporates alternative strategies and forms of testing	L	N	N	N	P

N = No coverage

L = Limited coverage: Task does not assess most elements of the framework with much depth or breadth

P = Partial coverage: Some dimensions of this element are covered

F = Multiple dimensions of the element are addressed in some depth

BEST COPY AVAILABLE

which teachers are held accountable for any element of the framework, what gets covered depends greatly on the specific tasks teachers complete and their scoring criteria. During piloting, each teacher only completed three of the six tasks. Also, scoring criteria change each time a new task is generated from a "shell" (see **Development**); therefore, coverage of various elements of the framework within tasks may vary from one test administration to another. Finally, readers should keep in mind that the simulation tasks were intended to be used in conjunction with other forms of assessment, which presumably would extend coverage of the *English/Language Arts Framework*.

Structuring a literature-based curriculum. The language arts framework places heavy emphasis on acquainting students with "significant literary works." Developers of the simulation tasks intentionally avoided constructing activities that necessitated content knowledge in order to be successfully completed. As a result, familiarity with specific literary works or categories of literature are not directly assessed. However, not being familiar with a specific literary work may have influenced some teachers' responses and the scoring of one of the scenarios in **Responding to Typical Problems**. In the scenario, teachers are to envision themselves taking over a class mid-semester that had begun reading Conrad's *Heart of Darkness*. The essence of the problem is whether to continue reading or abandon the text. The scoring was largely based on teachers knowing that the appropriate course of action for the scenario was to abandon the text students had begun with a previous teacher. One scorer, who was also a test developer, noted that if the text in question were less complex, it would not necessarily be as appropriate to shift to a new novel. In this instance, knowledge of the text was directly relevant to teachers' scores. Other scenarios in **Responding to Typical Problems** relevant to structuring a literature-based curriculum assessed teachers' ability to justify teaching a piece of controversial literature and recognizing the importance of looking at literature from different perspectives.

Scoring well on the **Designing a Lesson Sequence** task depended on teachers' ability to design lessons that incorporate "theme," an integral element of a literature-based curriculum. In both **Oral Presentation Skills** and **Stages of the Writing Process**, points are awarded if teachers recognize that an assignment is based on a piece of literature.

Implementing a meaning-centered curriculum that promotes critical thinking and integrates skills of reading, writing, listening and speaking. Another strong emphasis of the *English/Language Arts Framework* is that teachers avoid focusing on only one language art at a time (e.g., reading), without integrating it with other skills (such as writing, listening, and speaking). The thrust of this element of the framework is that students should develop knowledge in context, rather than in discrete bits. A separate but closely-related facet of this point is that developing critical thinking should be a cornerstone of teachers' curricula.

Designing a Lesson Sequence addressed this element of the framework most directly. Scoring criteria specified the "integration of all language arts" in teachers' lesson plans. Teachers were expected to include writing, speaking, reading, listening, and critical thinking activities throughout their week-long curricula. Parts of **Responding to Student Writing** assessed whether teachers made comments on student papers that focused on ideas rather than grammar or superficial details of style. Similarly, two of the **Responding to Typical Problems** scenarios required teachers to justify a focus on ideas rather than grammar in early drafts of student papers. Finally, in both **Stages of the Writing Process** and **Oral Presentation Skills**, teachers were awarded points for recognizing strengths in assignments that promoted both critical thinking and integration of various language arts.

Attending to various "stages" of the writing process. In recent years, composition research and the Bay Area Writing Project (now expanded into the National Writing Project) have greatly influenced curriculum development and instruction throughout the state. The *English/Language Arts Framework* explicitly states that features of an effective English/Language Arts curriculum will include "a writing program that includes attention to the various stages of the writing process--from prewriting through postwriting and from fluency and content through form and correctness." This element of the framework is one of the more thoroughly covered across the five tasks. Success in two tasks--**Responding to Student Writing** and **Stages of the Writing Process**--depends upon teachers' ability to apply principles of the writing process as outlined in the framework. In addition, two parallel scenarios in **Responding to Typical Problems** require teachers to justify focusing on ideas rather than grammar in early drafts of student papers (these scenarios necessarily cover both attending to stages of the writing process and promoting critical thinking).

Sensitizing students to values in literature that "reflect real dilemmas faced by all human beings...." This element of the framework focuses on using literature to directly address pertinent social and personal issues. One of the **Responding to Typical Problems** scenarios assesses teachers' ability to argue for the value of studying controversial themes in literature. Success in **Designing a Lesson Sequence** depends on teachers integrating the themes of the readings provided--which include issues of divorce--throughout the week. In scoring teachers for focusing on ideas rather than grammar in initial drafts of *Huckleberry Finn* papers, **Responding to Student Writing** also seeks teachers' adherence to this element of the framework.

Designing instructional programs that provide opportunities for critical thinking among all students, regardless of prior academic preparation or language ability. The influx into schools of students with minimal English literacy skills has influenced instruction in almost every district in the state. Among the fifty-six teachers who completed these tasks, for example, not one worked in a school in which English was the only language spoken; the mean number of languages spoken was five. The language arts framework emphasizes utilizing diverse instructional strategies to engage critical thinking in all students, not just those who are English proficient.

Every task addresses issues of limited-English speakers or those with less developed literacy skills. Two of the **Responding to Typical Problems** situations revolve around needs of limited-English speakers. One of the criteria for scoring **Designing a Lesson Sequence** has to do with how well teachers provide activities for a broad range of students' language and performance abilities. Two of the student essays presented in **Responding to Student Writing** incorporate errors obviously made by limited English proficiency writers. One part of the **Stages of the Writing Process** task has a provision to award points if teachers note that "All students are involved"; respondents earned points for articulating the value of group work for limited-English and low-ability writers. Finally, all three parts of **Oral Presentation Skills** award or deduct points for teachers who correctly or incorrectly identify facets of instruction pertinent to limited-English speakers or those of diverse ethnic backgrounds.

Developing oral language skills among all students. A feature of ineffective language arts curricula includes an oral language program "in which only the most verbally skilled students speak often or in which speaking is isolated from other language activities, such as reading and writing." Teachers' knowledge of oral presentation skills is directly addressed in the **Oral Presentation Skills** task. In **Oral Presentation Skills** sub-tasks, teachers identify strengths and weaknesses of assignments revolving around an oral presentation lesson. Limited attention is also given oral skills development in **Designing a Lesson Sequence**, where teachers earn points for including oral activities such as discussions or group presentations in their week's lesson plans. Teachers are also rewarded in **Stages of the Writing Process** for noting the value of incorporating a peer feedback session as a means of fostering oral skills development.

Contributing to a school environment where students are encouraged to read widely, write frequently, and listen and speak effectively. This particular element of the framework is addressed in a limited way by two parts of current tasks. In **Designing a Lesson Sequence**, teachers earn points for providing opportunities for students to bring to bear personal experiences to the themes being explored. Teachers who made links to popular literature or students' home environments were rewarded when it appeared they did so in order to promote broad notions of literacy. More directly, one of the scenarios in **Responding to Typical Problems** requires teachers to articulate the value of reading topics on a broad range of themes. Teachers were penalized when they suggested alternatives that failed to preserve students' independence in choosing books to read.

Modeling effective use of all language arts skills. The framework developers recognized the importance of good modeling for promoting literacy skills--not just by English teachers, but among all adults in schools. This element of the language arts framework was not addressed in any of the tasks developed for piloting.

Devising an assessment program that encompasses the full range of language arts goals and that incorporates alternative strategies and forms of testing. The intentions behind this element of the framework are two-fold. One is to be sure that students with weak English skills are not consistently failed because of tests that depend on language proficiency. Second, the framework developers recognize that all learners have preferred modes of internalizing knowledge. The *English/Language Arts Framework* encourages teachers to develop assessment strategies that provide opportunities for students to display diverse language arts skills. The framework also encourages teachers to devise assessments other than objective multiple-choice tests.

In **Responding to Typical Problem Situations**, there is a scenario that provides opportunities for teachers to suggest a creative assessment of students' progress through a text taught by a previous teacher; however, though teachers would be given a point for construing a creative and fair assessment, that is not the thrust of the scoring. Knowledge about alternative assessment is more directly tapped in **Oral Presentation Skills**, where teachers critique the structure of a (graded) oral presentation assignment.

Taken together, the simulation tasks developed cover the *English/Language Arts Framework* fairly thoroughly (see Table 6.2). Each of the five tasks covers at least two of the elements in the *English/Language Arts Framework*. One element of the framework, *modelling effective use of all language arts skills*, was not addressed at all.

A combination of decisions among policy makers and test designers will be needed to determine whether it is appropriate for all facets of the language arts framework to be covered each administration. Again, tasks can be modified to include a broad range of language arts objectives, but not without some sacrifice to test reliability. Furthermore, coverage of the framework for any one candidate necessarily drops as he or she completes fewer tasks, as was the case in this pilot testing. Finally, the extent to which teachers are ultimately held accountable for any element of the language arts framework depends greatly on the scoring criteria for a particular task, and the extent to which it is interpreted or adhered to during the scoring process (see **Performance on Structured Simulation Tasks**).

Extent of Coverage of California Standards for Beginning Teachers

The California Beginning Teacher Standards are criteria for teacher competence and performance that the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. Listed below are brief italicized descriptions of Standards 22 through 32, which pertain to expectations of student competencies to be attained prior to graduation from teacher preparation programs. (The remaining standards address programmatic requirements.) To evaluate the five simulation tasks and make inferences about their appropriateness for use with California secondary level English teachers, the stimulus materials and scoring criteria for each task were compared with the eleven California Beginning Teacher Standards (see Table 6.3 for a summary of coverage). Each standard and its coverage is discussed separately.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* It is difficult to conceive of a paper and pencil assessment that could assess a candidate's performance in this area. However, aspects of most tasks do tap teachers' knowledge of basic principles necessary for maintaining an equitable classroom environment. Three of the eleven **Responding to Typical Problems** scenarios focus on equitable treatment of students. Part II of **Oral Presentation Skills** provides numerous examples of a classroom teacher failing to respect and deal equitably with students. Teachers recognizing this in their answers heavily influenced their scores on that exercise. Scoring of the **Responding to Student Writing** task depended, in part, on the tone of teachers' comments on student papers, or on recognizing harsh comments written on student papers by hypothetical teachers. In **Stages of the Writing Process**, teachers were awarded points for recognizing the inequity of only having the best student papers read aloud in class.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* These skills lie at the heart of **Designing a Lesson Sequence**, where teachers are

TABLE 6.3

CONGRUENCE WITH CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Element of the Framework	Responding to Typical Problems A and B	Responding to Student Writing	Designing a Lesson Sequence	Stages of the Writing Process	Developing Oral Presentation Skills
22. Student Rapport and Classroom Environment	P	P	N	L	P
23. Curriculum and Instructional Planning Skills	L	N	F	P	P
24. Diverse and Appropriate Teaching	P	P	P	N	L
25. Student Motivation, Involvement and Conduct	F	L	L	P	P
26. Presentation Skills	N	L	N	L	L
27. Student Diagnosis, Achievement and Evaluation	L	P	N	L	L
28. Cognitive Outcomes of Teaching	N	N	N	N	L
29. Affective Outcomes of Teaching	L	F	P	L	P
30. Capacity to Teach Cross-Culturally	L	L	L	N	L
31. Readiness for Diverse Responsibilities	P	L	P	N	L
32. Professional Obligations	F	N	N	N	N

N = No coverage

L = Limited coverage: Task does not assess most elements of the framework with much depth or breadth

P = Partial coverage: Some dimensions of this element are covered

F = Multiple dimensions of the element are addressed in some depth required to create a five-day lesson plan revolving around core objectives, designing activities aligned with those objectives, and gearing instruction to appropriate grade and ability levels of students. Other tasks, such as **Oral Presentation Skills** and **Stages of the Writing Process** award points for teachers' recognizing the strength or weakness of activities, given the objectives presented.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities, and material that are appropriate for students with diverse needs, interests, and learning styles.* Some facets of each task require teachers to modify instruction for limited English speaking students. Two of the **Responding to Typical Problems** scenarios revolve around needs of limited-English speakers. One of the criteria for scoring **Designing a Lesson Sequence** is concerned with how well teachers provide activities for a broad range of students' language and performance abilities. Two of the student essays presented in **Responding to Student Writing** incorporate errors obviously made by limited-English proficiency writers. In one part of the **Stages of the Writing Process** task, respondents earned points for articulating the value of group work for limited-English and low-ability writers. Finally, all three sub-tasks of **Oral Presentation Skills** award or deduct points for correctly or incorrectly identifying facets of instruction pertinent to limited-English speakers or those of diverse ethnic backgrounds.

The five tasks developed less directly address differences students might have in ability other than language proficiency, nor do any tasks address the needs of students with diverse ethnic backgrounds. Regarding this last point, members of the development team said that, although they tried weaving into tasks examples of appropriate or inappropriate practice related to working with minority students, it was difficult to do without stereotyping.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* This standard is covered relatively extensively across the five tasks developed. **Oral Presentation Skills** asks teachers to identify strengths and weaknesses of lessons and transcripts reflecting appropriate and inappropriate motivational and discipline techniques. In **Responding to Student Writing**, teachers' tone, which is related to student motivation, influenced scorers' holistic assessments. In the **Stages of the Writing Process** tasks, teachers were awarded points if they correctly identified motivational strengths or weaknesses of assignments presented. In **Designing a Lesson Sequence**, scorers determined that barely acceptable lesson plans need not be particularly stimulating for students; however, lesson plans that emphasized student engagement were rated higher than those that didn't. Three scenarios in **Responding to Typical Problems** all focus to varying degrees on motivation of classes of students with varying abilities.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* This standard is not directly addressed by any of the tasks. However, teachers are asked to identify unclear instructions or assignments presented by hypothetical teachers in **Developing Oral Presentation Skills**, **Stages of the Writing Process**, and **Responding to Student Writing**.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* This standard is assessed in at least a limited way in four of the structured simulation tasks. **Responding to Student Writing** provides teachers with opportunities to diagnose strengths and weakness in student essays. Brief (one sentence) backgrounds are presented about each student author to help teachers evaluate papers in the context of the students' prior achievement. One of the **Responding to Typical Problems** asks teachers how they would take over a class in the middle of a term, based on what students had previously accomplished. Points were awarded when teachers discussed how they would evaluate students' progress to date, but diagnosis of students' progress was not a primary issue in scoring the responses. In **Oral Presentation Skills**, teachers comment upon a hypothetical teacher's evaluations of students' oral reports. Teachers are awarded points for correctly identifying strengths and weaknesses of the evaluations; however, this facet of scoring addresses the standard in only a limited way.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* The only task in which this standard is addressed is **Oral Presentation Skills**. In the transcripts presented, a hypothetical teacher makes comments about the quality of students' oral presentations, which teachers can identify as strong or weak. However, the instructions ask teachers to explicitly comment on only the hypothetical teacher's comments, and not directly about what students do or say. So, teachers' ability to recognize and remediate students' critical thinking skills is addressed only indirectly and in a limited way.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward subjects learned, the students themselves, and their capacity to become independent learners.* This standard is addressed in each of the five structured simulation tasks. In scoring **Responding to Student Writing**, scorers were greatly influenced by how a teacher's response would affect a student's motivation to write. This concern was weighed about equally with concerns for correctly identifying analytical flaws. In the portion of the **Responding to Student Writing Task** where teachers respond to comments made on student papers by a hypothetical teacher, successful scores on two of the three examples depended on teachers recognizing the inappropriately harsh tone of comments. Points could be earned on the **Oral Presentation Skills** task for correctly identifying how a hypothetical teacher's comments would motivate or discourage students from studying literature or participating in future oral reports. In **Designing a Lesson Sequence**, scorers assessed how well teachers linked the literature studied to the lives of students. It is generally recognized that such connections make learning meaningful, thus motivating students to participate and encouraging them to extend their knowledge on their own.

Standard 30: Capacity to Teach Cross-Culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic, and socioeconomic differences.* Contextual elements are built into all tasks, which specify students' ethnic and or linguistic background, as well as gender. Usually the instructions specify a "heterogeneous classroom," which presumably would include students different from the candidate. Nevertheless, assessment of this standard depends on how it is built into the scoring criteria, the background of the teachers and scorers, and how different perspectives are incorporated into the scoring process. On one hand, a paper and pencil simulation may not be the best means of assessing teachers' cross-cultural competence. To the extent that this standard might have been attended to, the scoring criteria and scoring process did not adequately address these issues. Most scorers were Anglo, with no demonstrated competence dealing with students different from themselves. In one case, a teacher experienced working with limited-English speakers was deferred to by the scoring team evaluating responses to a task involving ESL instruction. Although the scoring team was probably prudent in deferring to the expert, in a case such as ESL instruction, where professionals throughout the field greatly differ about what constitutes effective practice, it is questionable whether one scorer's opinion ought to define what constitutes an acceptable response. (For a fuller discussion of these issues, see **Recommendations**.)

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* All tasks include classroom context as one of the variables to take into account while formulating responses. Most activities are set in 10th- through 12th-grade English classes.

Teachers taking the test did not seem to perform better or worse as a result of their experience being mostly with high school or junior high students. Ability differences are mostly presented in terms of language proficiency; however, teachers are rewarded in **Designing a Lesson Sequence**, **Developing Oral Presentation Skills**, **Stages of the Writing Process**, **Responding to Student Writing**, and some scenarios in **Responding to Typical Problems** when they identify situations where needs of lower-ability students are not accommodated.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interaction with other members of the profession.* Six of the eleven scenarios in **Responding to Typical Problems** focus on professional conduct, either with students, parents, or other school personnel. Scoring heavily depended on the extent to which teachers recognized and respected the parameters of their professional responsibilities. Although it is difficult to tell the extent to which a teacher's knowledge translates into practice in such matters, the prompts seemed to elicit a good sample of a teacher's knowledge and attitudes regarding professional conduct. Other tasks did not focus on this standard; however, when teachers' responses reflected poor ethical judgement, points were deducted.

Table 6.3 lists the beginning teacher standards and an evaluation of the extent to which each is covered across all tasks. As with coverage of the *English/Language Arts Framework*, the extent to which beginning teacher standards are covered will ultimately be influenced by the following four factors:

- the number and type of tasks candidates complete;
- the skills and knowledge assessed each time a new task is developed;
- policy decisions during task development regarding the extent to which reliable assessment of each standard will be sacrificed to breadth of coverage across all standards; and
- the extent to which scoring guides created during task development are adhered to during the scoring process.

The previous two sections have addressed the coverage of each of the tasks with state standards. To address other aspects of the tasks' content, teachers and scorers completed evaluations of the tasks regarding a number of dimensions: job-relatedness, appropriateness for beginning teachers, appropriateness across teaching contexts, fairness across groups of teachers, and, finally, a general evaluation of the appropriateness of this method of assessment. Their perceptions, as well as a discussion of the scoring process and results, are included in the remainder of this section.

Job-relatedness

Both teachers and scorers were asked whether the tasks chosen were relevant to the job of teaching English at the secondary level.

Teacher perceptions. Teachers were asked: Do you feel the tasks chosen for this assessment are relevant to your job of teaching secondary English? Ninety percent of teachers in both Group I and Group II responded "Yes," with ten percent responding "No" or "Yes & No." Among Group I teachers who elaborated their response, most focused on the **Designing A Lesson Sequence** task. Some teachers felt that that task was the most relevant of the three they completed. In fact, a few teachers contacted FWL assessors requesting their lesson plans so they could actually use them in their classes. Others felt that having two hours (the allotted time limit) to design a five-day lesson plan was a luxury not normally experienced in actual practice. One teacher wrote:

> *Yes, I comment on papers, structure lessons, and handle problems, but I certainly don't have the luxury of the extended time limit for each activity.*

This same teacher suggested that a 15-20 minute sharing of ideas should be built into **Designing a Lesson Sequence**, emphasizing that many lessons result from brainstorming with others and that creating a new lesson each day leads to burnout.

Two teachers in Group I felt that the tasks were not relevant because they hadn't taught at the grade levels specified in most tasks (grades 9-12). One teacher noted that the tasks were relevant only "[i]f you want to know what I *think*, not what I actually *do*." Another added, "What a person may write, however, may not be the way they perform in a classroom." A Group II teacher felt that the parts of **Oral Presentation Skills** and **Stages of the Writing Process** that dealt with assessing the structure of lesson plans and revising them were relevant, but that the rest "did not seem to be of much worth."

Scorer perceptions. All scorers believed that structured simulation tasks were relevant to beginning secondary level English teachers. A few scorers believed that some tasks ought to be weighted more heavily than others, presumably because they capture classroom-related behaviors more strongly. **Designing a Lesson Sequence, Responding to Student Writing** and **Responding to Typical Problem Situations** were those listed as deserving more weight.

Appropriateness for Beginning Teachers

Teachers were asked the following question on the evaluation questionnaire: Do you think this type of assessment (i.e., a set of structured simulation tasks) is an appropriate way of assessing your competency in teaching secondary English? Teachers' reactions were mixed. In Group I, 64% responded "Yes," 25% responded "No," and the rest, 11%, responded "Yes and No." In Group II, 50% responded "Yes," 36% responded "No," and 14% responded "Yes and No." In reality, these percentages are not precise because of qualifications many teachers included in their responses. Of the 19 teachers (of 28 Group I teachers altogether) who elaborated on their answers, only 5 did so unequivocally. Teachers expressed concerns about the subjectivity of scoring, narrowness of assessment, and the artificiality of tasks compared to actual practice. Despite these concerns, teachers were encouraged about the general nature of structured simulation tasks, as exemplified in the following comments:

> *This focuses on* ***what*** *English teachers do and what we should be doing--* ***teaching.***
>
> *Being able to* ***write*** *responses rather than multiple choice is a free, accurate way to respond. (Especially English teachers, since writing* ***should*** *be their strongest asset!) The simulations also prompt thoughts on preparation and reaction. I found I learned a lot, simply by being asked to respond and create from my own mind.*
>
> *Giving descriptions of students as variables to how things were done was helpful--nothing is cut and dry in your lessons.*

One respondent, who also participated in a pilot test of another assessment, the Secondary English Assessment: Performance-based Exercises, developed by San Francisco State University (see Chapter 7), said:

> *This assessment, in comparison to the one I did last summer, was much better. It seemed to be a more appropriate way to assess people. The oral component to the first assessment would seem more difficult to judge. The week-long lesson planning [**Designing a Lesson Sequence**] was more specific (and took much less time) than the unit planning I did for the first assessment.*

Several teachers who believed the simulations were appropriate also expressed the hope that these tasks would be used in conjunction with other means of assessments to determine overall teachers' competence. Test developers share this hope.

Four teachers discussed concerns they had about the **Responding to Typical Problems** task:

> *I suppose that more hypothetical situations would help the "[**Responding to Typical Problems**]" section. More of those related to possible areas of censorship, classroom management techniques, applying to all teachers, etc. might be appropriate.*
>
> *[Tasks] don't allow for vast differences in teaching assignments and in teacher conduct. Some of the situations given in [**Responding to Typical Problems**] would never arise in my classroom because I wouldn't allow it, or it is contrary to my teaching style.*
>
> *...my concern is the subjectivity of grading. There are many ways to handle a situation and more than one way can work!*
>
> *What we do on paper is not, at times, a clear indicator of what we do in class. Intellectually, you may know the "proper" way to handle a situation, but your on-the-spot reactions to daily situations might vary drastically.*

The predominant complaint among those who felt that the tasks were an inappropriate way to assess their competence was that performance on the tasks reflected knowledge more than ability. One respondent's answer epitomizes that of many:

> *Just because someone can answer these types of tests does not mean that they will be a competent teacher. I may know the "correct answer,"*

> *but if I can't relate to my students or get the information across to them, I certainly cannot be considered competent.*

Two teachers emphasized how not being able to display their ability working with real students makes the tasks artificial:

> *It was a false setting (no students, room, supplies, etc.). Much of my teaching style has to do with the way I interact with the kids.*
>
> *This test strips away the single most crucial aspect that contributes to teacher success and effectiveness--and that is the teacher/student bond.*

Finally, two teachers noted that although the tasks were good for "assessing teaching skills," they were not necessarily good for "the teaching of English."

In addition to commenting about the appropriateness of the assessment, teachers were asked whether the tasks were too easy or too hard. Although only 32% (9 of 28) Group II respondents said that the assessment was too easy, each of these teachers elaborated a response, and each of their responses focused on the **Oral Presentation Skills** exercise. As the following comments suggest, weaknesses of the hypothetical teacher were far too obvious:

> *The response to the [hypothetical] teacher [portrayed in the **Oral Presentation Skills** script] was almost an insult. Her responses and comments were so stupid that any fool could catch her errors, especially after all of those great psychology courses we have to take while getting our credential.*
>
> *Identifying the weaknesses in the "script" was truly ridiculous--anyone can respond to the obvious teacher failings it contained, while again not revealing anything about what that person's teaching method and style would be like.*

In Group I, 18% (5 of 28) respondents who noted that tasks were easy referred to **Responding to Typical Problems**. One teacher felt that the scenarios were "caricatures," too obvious. Another teacher felt that the task may have seemed easy because it reflected situations that he constantly deals with in his teaching. Other teachers did not elaborate their "Yes" or "No" answer.

Eighty-six percent of the teachers in both groups felt that the tasks were challenging, but not too difficult. Time constraints pressured some respondents, but the tasks themselves seemed reasonable. Among the elaborated responses, **Designing a Lesson Sequence** was mentioned most often as being difficult due to (1) the format in which the lessons had to be written (i.e., objectives, activities, rationale, and time allotments), (2) respondents being tied to having to plan for exactly five days (and not more), or (3) (in one case) a respondent's feeling that more room was needed on the response sheet. Group II respondents who felt the tasks were difficult focused on time constraints, noting that the exercises were challenging and therefore exhausting to complete in one sitting.

Teachers were also asked which parts of the assessment, if any, could be given after student teaching but before teaching a classroom on their own. Teachers differed as to how much classroom experience would be necessary in order to perform well on any or all tasks. Teachers were about equally divided about which tasks, if any, could be successfully completed after student teaching. Many respondents suggested specific tasks that should be mastered right after student teaching, but each of the five tasks was mentioned at least once across all respondents. Only three of fifty-six respondents reported feeling unprepared by both their teacher education programs and classroom experience to complete the tasks at all; however, they did not elaborate their responses.

Scorer perceptions. All scorers agreed that the simulations tapped knowledge that beginning teachers should have. In those tasks where teachers' performance as a group was low, scorers believed that teacher education programs, not the assessment, were at fault. Regarding **Designing a Lesson Sequence**, one scorer said:

> *The [California] state [language arts] framework opens the door for teachers to be their own curriculum developers more than ever before. Are teacher education programs preparing teachers to do that? Since we have a batch of candidates who can't do this, you wonder if they're being prepared adequately.*

Another scorer commented:

> ***Responding to Student Writing** is particularly good. I suspect [teachers] lack training more in responding to student writing than in other areas...Because the [student author of the essay teachers marked] was LEP, it might show the ignorance of teachers in assessing problems in writing when a student is ESL.*

Overall, scorers felt that the simulation tasks could be used to define standards for teacher education programs.

As far as appropriateness of specific tasks, no scorers identified any one task as inappropriate. All scorers responded on evaluation questionnaires that the tasks and questions were appropriate means of assessing secondary English teachers' pedagogical and English teaching skills; however, some did make comments about the content within certain tasks.

Two scorers commented upon the obviousness of the hypothetical teacher's deficiencies in **Developing Oral Presentation Skills**. One scorer felt that the task would better tap teachers' knowledge if teachers confined their observations to identifying the hypothetical teacher's strengths, rather than both strengths and weaknesses. This particular scorer was particularly concerned about beginning teachers evaluating other teachers. First, new teachers may not have the skills to evaluate other teachers fairly; institutionalizing such evaluation by this kind of assessment could promote an unhealthy practice among new professionals, who tend to be overly critical anyway. Second, using obviously poor models of teaching throughout the assessment can leave test-takers with a distorted view of other teachers, and more importantly, themselves:

> *I felt that [teachers'] answers showed a certain smugness and showing off. I think these people left the test feeling that they were above the sample "teachers" in the test. I hope this was just my impression, not a reality. Teaching can be improved constantly. I would hope that those who successfully completed the test would not feel that they had mastered teaching.*

One scorer, a classroom teacher and department head, commented about the appropriateness of **Responding to Student Writing** for *all* beginning teachers, not just English teachers:

> *...Writing is not the exclusive responsibility of English teachers. All teachers should be accountable for this information if we're going to raise standards across the board.*

Overall, scorers were enthusiastic about the five tasks and felt that, with modifications to some scoring guides or procedures, they covered important knowledge beginning English teachers should have.

Performance on Structured Simulation Tasks

Conceiving teachers' scores as a reflection of their ability is misleading if one assumes *a priori* that the instrument is valid and the scoring process was reliable. Any interpretation of teachers' scores must be understood in terms of the process used to obtain them. In this section, scores obtained on the structured simulation tasks are analyzed in terms of the reliability of the scoring process used to obtain them. A later section, **Validity**, discusses the instrument's validity.

The scoring process that transpired should be understood as a necessary stage of the development process, not as the end result. As the following sections will indicate, structuring and monitoring of the scoring process presents one of the greatest challenges to future development of the simulation tasks. Any further development of these structured simulation tasks must attend to issues common to development of these types of assessments, as well as to obstacles unique to these particular tasks.

A discussion of actual scoring procedures for each task and sub-task will follow. First, however, we provide a general model of the scoring procedure. The usual procedure for scoring tasks and sub-tasks was as follows:

1. The RAND developer on each team provided a brief overview of the task, then instructed raters to read the stimulus materials and scoring guides.

2. The RAND developer then asked whether there were any questions. If so, they were answered.

3. Raters were instructed to score one or more responses, using the scoring guide. Initially, there were usually notable discrepancies among scores based on differing interpretations of criteria, interpretations of responses, and/or differing attention or weighing given to various aspects of the responses.

4. Through discussion of criteria and interpretation of responses, consensus could generally be reached among group members within the first ten papers. That is to say, scores would begin to coincide more often among raters. Further discrepancies would be worked out among members of the scoring team on a case by case basis.

It is important to understand that in all cases scores were arrived upon by consensus. The process used to reach consensus depended heavily on interpersonal dynamics. In those initial discussions in which the bases for consensus were developed, individuals whose interpersonal styles were more dynamic or aggressive tended to overwhelm those with more subtle interpersonal styles. At various times, members of scoring teams were interrupted, argued down, or ignored. Some scoring team members' opinions tended to be valued more than others, often for reasons other than expertise. In addition, some who had less experience scoring were less confident of their own abilities, especially when just beginning. Such scorers tended to remain silent more often, even when they had expertise that could have contributed to scoring decisions. These dynamics were unintentional and often unnoticed by scorers. Nevertheless, the effect tended to be that consensus resulted not always from agreement, but from abdication to a dominant personality. Observing this dynamic across scoring teams leads to the strong recommendation that if consensus is used as a scoring technique in the future that it be implemented only with a carefully designed protocol that is rigorously followed (see **Recommendations**).

The range of scores for various tasks and sub-tasks varied greatly. Because some sub-tasks were scored analytically and some holistically, correspondence between scores on different sub-tasks is minimal. In other words, a score of 3 on one sub-task can mean something completely different than the same score on another sub-task; two sub-tasks having the same range of scores do not necessarily have the same range of quality.

To help interpret the data, we converted the scores for each sub-task to pass/fail. For most sub-tasks, this conversion was based on the scorers' judgement of the level of performance which represented minimal acceptable competence and that which did not. For other sub-tasks, the pass/fail cut-off score was determined by FWL staff listening to informal discussions held during the scoring process and evaluating the pertinent scoring criteria. It should be noted that the developers of the simulation tasks intentionally did not build pass/fail criteria into the scoring framework because such decisions are usually considered matters of policy.

The range of raw scores for each sub-task, their mean, and the percentage of teachers who would have passed based on score conversions to pass/fail are listed in Tables 6.4 and 6.5.

TABLE 6.4

RANGE OF SCORES, MEAN, AND PASS RATES FOR TASKS FOR GROUP I

Task	Range	Median	Percentage of Teachers Passing*
Responding to Typical Problems – Form A			
#1	–2 to +2	0	18%
#2	–1 to +4	2	50%
#3	–1 to +3	0	32%
#4	–1 to +2	2	72%
#5	1 to 5	2	32%
#6	–1 to +3	2	86%
Responding to Student Writing			
Part I:			
Essay #1	1 to 6	2	25%
Essay #2	1 to 5	3	61%
Part II:			
Essay #3	–1 to 6	2	25%
Essay #4	–2 to 7	0	39%
Essay #5	2 to 9	3	79%
Designing a Lesson Sequence	1 to 5	2	36%

*Cut off scores determined using scoring criteria and expert judgement of scorers and/or evaluator.

TABLE 6.5

RANGE OF SCORES, MEAN, AND PASS RATES FOR TASKS FOR GROUP II

Task	Range	Median	Percentage of Teachers Passing*
Responding to Typical Problems - Form B			
#1	-1 to +2	-1	36%
#2	1 to 5	2	39%
#3	1 to 5	2	25%
#4	1 to 4	1	36%
#5	1 to 4	2	75%
#6	1 to 3	2	100%
Stages of the Writing Process			
Part I:			
Assignment	0 to 6	3	82%
Activity 1	1 to 5	2	82%
Activity 2	1 to 5	2	39%
Activity 3	1 to 3	2	71%
Activity 4	1 to 4	2	36%
Activity 5	1 to 3	1	46%
Activity 6	1 to 2	1	46%
Activity 7	1 to 4	1	36%
Activity 8	1 to 5	3	54%
Part II	1 to 5	2	68%
Developing Oral Presentation Skills			
Part I	0 to 9	5	78%
Part II			
Segment 1	2 to 15	10	50%
Segment 2	3 to 24	14	78%

*Cut off scores determined using scoring criteria and expert judgement of scorers and/or evaluator.

The following sections review the scoring process in detail, including discussions of analytic versus holistic scoring approaches, issues arising in scoring of specific tasks and sub-tasks, and setting of standards for acceptable performance.

Analytic versus holistic scoring. The analytic and holistic scoring methods used to derive scores are discussed earlier (see **Scoring Process**). The greatest difference between analytic and holistic scoring was that, in holistic scoring, specific elements of a response had the potential to influence the overall assessment of that response much more than if an analytic method were used. For example, in **Responding to Typical Problems**, one scenario asked teachers to determine how they would handle an ESL student, new to a mainstream class, who expressed concern about an upcoming in-class writing assignment. One teacher responded:

> *I would allow Tuan to present his assignment orally. Tuan would be given a tape recorder and some privacy, then he could verbalize what he had written in his journal. Later, Tuan and I could transcribe the recording so he could see a finished product. In time, this might alleviate Tuan's fear of writing by showing him that it is really just another form of talking.*

The scoring team using the holistic method reacted harshly to the last sentence, protesting that speaking is indeed *not* just another form of writing. Though they admitted that the strategy might have some merit, the teacher's underlying belief about the relationship between writing and speaking, as expressed in the last sentence, was clearly unacceptable. Because of the rationale elaborated in the last line, this response was given the lowest rating. The re-scoring team, however, using an analytical method, responded to the merit in the teacher's approach. Three points were awarded for positive elements of the response, and one deducted for the mistaken last sentence. One scorer commented, "This is exactly what Steve Krashen has been preaching all over town!" (In fact, Krashen is listed in the bibliography and recommended reading list of the *English/Language Arts Framework*.) In the end, this response was rated among the highest of those re-scored.

Another example from **Responding to Student Writing** also illustrates the point. In Part I, where teachers wrote responses to student essays, one teacher's handwriting was especially large. Though her comments were considered accurate, scorers believed that a student faced with a paper as full of teacher marks as this one appeared to be would be daunted by the prospect of revision--even before reading the comments. As a result, the paper was placed in a low category. Were analytical scoring used, the appearance of the paper might have resulted in a point deduction, but probably would not have leveraged the response so dramatically.

Differences between analytic and holistic scoring methods raise issues pertinent to test reliability. Though holistic scoring seems more intuitively valid, holistic scores can be difficult to replicate since factors that sway holistic ratings vary from one professional to another and one scoring team to another. Such problems are minimized when the assessment objectives are clear for each task and the criteria for scoring clearly elaborated, which wasn't the case in these simulation tasks.

Concerns about reliability of holistic scoring surfaced most obviously in areas where common beliefs about what constitutes acceptable practice were hard to codify. In such areas as ESL instruction, curriculum planning, implementation of cooperative learning, and developing students' critical thinking skills, vehement though reasonable disagreement can exist among equally qualified professionals whose philosophies are shaped by their unique training and experiences. Such disagreement existed among the professionals who scored these structured simulation tasks. In fact, RAND developers commented that consensus about what constitutes acceptable English/Language Arts instruction was consistently more difficult to come by than was consensus about what constitutes acceptable life/general science instruction.

Holistic scoring also seemed to become less reliable as the criteria upon which decisions were reached increased. In the **Designing a Lesson Sequence** task, for example, scorers considered 18 criteria in deriving a single holistic score. Although both the original and re-scoring teams considered the same criteria, their ratings of responses differed in part because of differing emphasis on various criteria. Fewer criteria, or more scores based on smaller groups of criteria, might have improved the reliability of scores.

Ultimately, the inclination towards holistic scoring must be balanced against achieving reliable results, especially for high stakes, summative judgments, such as these tasks were designed to be.

Performance on Responding to Typical Problems. As shown on Tables 6.4 and 6.5, range of scores varied from one scenario to another. Form A (completed by Group I teachers) was scored analytically, with the exception of scenario #5, which was scored holistically. Form B (completed by Group II teachers) were, with the exception of the first scenario, scored holistically. Although Forms A and B were not completely parallel, four of the six scenarios embody the same issues across the two forms. Scenario #1 of both Form

A and Form B is identical, presenting a situation where a teacher has assigned *Heart of Darkness* to be read by students, but is then taken ill one month into the semester. Teachers are to put themselves in the position of replacing the teacher, having to make the decision whether to continue or abandon the text. Of Group I teachers, 18% scored in the range considered to be acceptable practice, compared with 36% in Group II. The two groups' scores were not distributed equally, with more teachers in Group II scoring toward the bottom range of the scale, while Group I teachers' scores are more evenly distributed.

Scenarios #2 and #5 on Form A and Form B, respectively, also parallel each other. These scenarios deal with confronting students who have cheated or plagiarized. Based on determinations of acceptable practice, 50% of Group I teachers and 75% of Group II teachers scored in acceptable ranges. Scenarios #4 for Form A and #2 for Form B were also parallel. At the heart of these two typical problem situations is a teachers' need to justify teaching a piece of literature to either students or parents. Of Group I teachers, 32% scored in a range of acceptable practice, compared to 39% of Group II teachers. Finally, scenario #6 in both forms A and B tapped teachers' ability to respond to parent complaints about their teaching of writing as a process, which necessarily de-emphasizes grammar and spelling in early stages of drafting. Eighty-six percent of Group I teachers and 100% of Group II teachers performed in the range considered to be acceptable practice.

Fourteen of twenty-eight responses to scenario #5 (Form A) were re-scored by an independent scoring team. The re-score team scored using the analytic method. Correlation between the original and re-scored scores was r=-.04, indicating poor reliability across the two scoring teams (see **Reliability**).

It is difficult to interpret the scores generated by teacher responses to the typical problem situations. Where scores are similar for paired scenarios, results could indicate in a broad way teachers' understanding of the issues involved. On the other hand, score distributions between parallel scenarios differ, indicating underlying differences between Group I and Group II teachers that may be the result of the different contexts in which they work, or some other, unaccounted for, variable. Poor reliability between two scoring teams suggests that scores are, in part, the result of different perceptions across scoring teams--perspectives which might or might not be replicable.

Performance on Designing A Lesson Sequence. This task was initially designed to generate five scores; however, because raters believed it more valid to judge the five-day lesson plans as a unit, only one score per teacher resulted. While scores were fairly evenly distributed, 36% of the teachers scored in the range considered to be minimally acceptable.

This task was also re-scored by an independent scoring team. Both teams used the same criteria, and both used a holistic scoring method. The original scoring team's range was from 1-5 (5=high); the re-score team's range was 1-7 (7=high). Correlation between the two teams was r=.28 on raw scores, r=.41 for scores adjusted to pass/not pass based on scoring team determinations of minimal acceptable competence.

In that lesson planning is expected of teachers in the course of their classroom practice, this task had much promise insofar as its ecological validity is concerned. However, difficulties arose in the scoring that would need to be attended to if such an assessment is to be used in the future.

For example, because the task is so open ended, a number of facets of lesson planning are able to be tapped--e.g., motivation, adhering to objectives, attending to diverse language and ability levels, integrating various aspects of language arts, to name a few. Having one score serve as a descriptor of one's ability in all of these important facets is a daunting challenge. In fact, combining the original scoring criteria with those scorers developed, raters had to attend to 18 scoring criteria in order to arrive at one score.

While qualitative differences between lesson plans were apparent, weighing of one criterion over another shifted over the course of the scoring sessions. For instance, early in the scoring, "integration of reading, writing, speaking, and listening" was considered extremely important, especially since such integration is a prime directive of the *English/Language Arts Framework*. Initially, ratings were heavily influenced by scorers' determinations about whether each of the language arts was focused on throughout the week's lessons. But over the course of the day and one-half in which responses were scored, emphasis on integration of all language arts seemed to shift. For example, toward the end of the scoring, one respondent was rated highly who did not emphasize reading--no provision was made in the lesson plans for students to read the pertinent literature on their own. Members of the scoring team were satisfied that this teacher's lesson plans should be rated highly because the teacher provided opportunities for students to respond to the literature pieces in group discussions. It is possible that if that particular response had been scored earlier in the process, it would have been rated differently.

What seems important about one lesson plan may not seem as important in another. Using 18 scoring criteria enabled scorers to justify the qualitative differences they did see; however, such an approach inevitably leads to decreased reliability. In fact, both scoring teams independently arrived upon the same scoring criteria. Differences in final estimates resulted less from differing interpretations of what respondents wrote than they did from differing emphases by scorers on various criteria during the scoring process.

Performance on Responding to Student Writing. Five scores were generated from this task, one for each student essay teachers commented on or about. Part I essays, where teachers commented on essays as if they were first drafts, were scored holistically. Part II essays, where teachers listed strengths and weaknesses of a hypothetical teacher's comments on three student essays, were scored analytically.

Overall quality of responses varied greatly from one essay to another, according to scorers. Based on the designated "minimal acceptable competence" cut-off for each essay, seven of 28 teachers performed adequately on the first essay in Part I, and 17 of 28 on the second. In Part II, 7, 11, and 22 respondents performed with at least minimal competence on each of the three remaining essays (n=28). Scores were generally not distributed evenly for each essay, although the quality of responses varied such that low scores on some essays still fell within a range of acceptable responses. In this task as others, interpretation of teachers' scores must be made in light of the scoring process.

Part I responses were considerably open-ended. This presented challenges to scorers. The challenge was heightened by the fact that this was the first task scored by the second-week scoring team, who were therefore new to the scoring process. As a result, problems arose in calibration among raters because initial teacher responses weren't fully discussed. After scorers read the instructions and materials, which included the scoring criteria, they were told to read five teacher responses. After that, the following discussion ensued (S refers to scorer, number designations refer to specific respondents):

S1: We seem to have a range of quality. How would you rate response X relative to Y or Z? Is there any that are worse than the others we've looked at? What about the scoring guide? We don't want to force the scoring guide [if it doesn't work for this essay], but let's look at it for the first few.

Let's look at T. It's clearly better than Y, but is it better than others?

S2: Response R has more specific references.

S1: Needs a definition of hero...?

S3: It's empathetic, the tone is supportive.

S2: How will it benefit the student, that's what I ask. So, how are we going to look at these? In terms of how well the teacher addresses these issues, or whether they do?

S1: Both. So we're saying that T is higher than R but less than X.

S3: Is B better or worse than X? It doesn't say anything positive, or about organization.

S2: S3's points are well-taken.

S1: So you'd put it worse than X? Let's try W.

In this interchange, the point that S2 raises regarding whether a rating is based on teachers addressing issues or addressing them well is pretty much lost in the discussion. The result was that S2 wasn't able to participate with confidence in the scoring process until she picked up on scoring guidelines intuitively. As scoring continued, initially reticent raters began asserting themselves into the consensus process, but not always. Often, they would change their rating without saying they had, when more dominant personalities announced their score first. As a result, it was not always clear even to raters why one essay was rated more highly than another. Toward the end of scoring for each of the Part I essays, consensus did seem to be reached more quickly and with greater conviction among all raters. However, the essential point to be made for future development is this: *in tasks with the greatest amount of open-endedness--especially if they are to be rated holistically--full discussion of rating criteria must occur before and during scoring so that it is explicit upon what bases rating decisions are made.*

Later, scorers noted that it was more difficult to calibrate answers in Part I of this task than it was in other tasks. Much of the problem centered on interpretations of teachers' comments on student papers. Teachers agreed, for instance, that a harsh tone in comments would be rated low. However, scorers differed in their perceptions of what constituted harshness.

In spite of the problems that arose during the scoring process of **Responding to Student Writing**, the ratings correlated more highly with each other overall than did intra-correlations of other tasks (most r's are in the .32 range; correlation between the two essay scores in Part I is r=.49 (r=.32 when scores are adjusted based on pass/not pass determinations). Statistical reliability was slightly higher than in other tasks (Cronbach alpha for standardized scores =.678.). Since the task corresponds closely with activities central to classroom practice, future focus on ironing out scoring difficulties could yield worthwhile benefits.

Performance on Stages of the Writing Process. This task generated ten scores. The range of quality and distribution of scores across sub-tasks varied considerably. The lowest percentages of minimum acceptable competence were 36% on teachers' critiques of Activity 4 (creating peer response groups to give feedback on first drafts) and Activity 7 (planning to grade final drafts of papers outside of class, commenting about their content and narrative quality). The highest percentage of minimum acceptable competence was 82% for both the teachers' ability to critique the initial assignment and Activity 1, the teacher's leading of a class discussion about the reading.

Correlations among the scores were low, the average being r=.16 (raw scores). This was surprising, considering that all sub-tasks 1) tested one's ability to recognize and apply knowledge about the writing process, and 2) were scored by the same scoring team. Scorers believed that, although teachers could articulate elements of the writing process, they were not well-trained to recognize and apply them. This is possible, but assumes *a priori* the viability of the assessment instrument. Nor does the scorers' explanation explain why teachers would perform so strikingly well and then strikingly poorly on similar tasks. For instance, in Activity #1, teachers evaluate strengths and weaknesses of class discussion as a pre-writing activity. Based on cut-off scores derived from scoring criteria, 82% (23 of 28) of the teachers responded adequately. On Activity #2, teachers evaluate strengths and weaknesses of another pre-writing activity based on an outlining technique that includes sharing story ideas in pairs. In this activity critique, only 39% (11 of 28) responded adequately, based on cut-off scores. It would not appear that the two sub-tasks assess significantly different kinds of knowledge; both sub-tasks deal with pre-writing activities.

Challenges faced in the holistic scoring of other tasks also surfaced during the scoring of **Stages of the Writing Process**. Raters were not always clear about the bases for rating judgments. One scorer simply stopped deducting points from responses because it was not clear to her on what grounds misidentifying a strength or weakness was ignored or penalized. In cases where the scoring guide was used, matching teacher responses to scoring criteria was not always consistent. One respondent, for example, was not given credit for writing "Adequate time is given for revision," even though the scoring guide awards a point for recognizing that "Time in class [is] given for revision." Another example has to do with a criterion that awards a point for recognizing as a weakness that "No discussion of revision [is provided by the teacher]." One candidate responded, "Previous writing should be put on overhead so students know what to shoot for." Though scorers interpreted this response as a discussion of revision, they also determined that acceptable practice would include discussion of revision within the context of the current assignment. Since the teacher suggested using a previous assignment, no point was awarded. However, a point was awarded to teachers who responded more vaguely (e.g., "More discussion should be given re: how to go about making revisions....").

Overall performance on **Stages of the Writing Process** was not strong, based on the scorers' ratings. However, more attention needs to be paid to the prompts to be sure they elicit the kinds of information scorers are seeking. Also, the scoring process should be revised so that the means of arriving upon scores are 1) clear to scorers, 2) based on generally agreed upon notions of acceptable practice, and 3) applied in as systematic a way as possible, given the difficulty of achieving highly reliable scores with such open-ended tasks.

Performance on Developing Oral Presentation Skills. Three scores were generated from this task, one for Part I, and two for Part II. Part I scores ranged from 0 to 9. Overall, respondents performed well on Part I of the task, based on the scoring criteria. Of the 27 responses to this part of the task, 21 were rated in the range of acceptable practice. Consensus was reached fairly early on in the scoring process and discrepancies were not great.

Part II included two sub-tasks, each of which required identification of the strengths and weaknesses of a hypothetical teacher's behavior and comments as depicted in a transcript of a class session. Scores for Part II ranged from 2 to 15 on the first sub-task, and 3 to 24 on the second sub-task. Although the two sub-tasks were very similar, only 50% of the teachers were judged as passing the first sub-task versus 78% for the second sub-task.

Correlation between the two segments of Part II was r=.12, which is surprisingly low considering similarities among them. However, the two segments were scored by different scoring teams, so the low correlation may be a reflection of poor inter-rater reliability. The scoring team that rated Part I also rated segment 1 of Part II. Correlation between those two sub-tasks is r=.32, which would be an expected correlation for two activities that measure different facets of a similar thing. Correlation between Part II, segment 2 and Part I (again, rated by different scoring teams) is r=-.04.

Nothing on the surface would suggest significant enough differences between the two segments that would explain their low correlation. It is the case that ten teachers ran out of time completing this task. Though their scores were not factored into the correlation calculations, time pressure might have influenced the responses of those who did finish. Yet, there is at least a 60% overlap of scoring criteria in the two parts, and that which doesn't overlap directly cannot be construed to be tapping a distinct enough knowledge base to warrant such low correlation between them. It is, therefore, a more likely possibility that low correlation between the two segments of Part II is a reflection of differing perspectives between scoring teams.

Determining performance standards. Making meaning of teachers' scores--on this or any other assessment instrument--depends upon how determinations of overall acceptable competence are made. This is a policy decision, not a psychometric one. If the goals of the assessment are to determine minimum acceptable competence, and each sub-task measures a different facet of teaching competence, then, ideally, one should pass each sub-task in order to pass the entire assessment. In fact, only one of fifty-six teachers "passed" all parts of the three tasks she completed, based on cut-off scores derived by scorers and/or the evaluator. Considering difficulties obtaining high reliability in scoring at this stage of development, it is surprising that even one teacher performed so consistently. Numerous techniques exist for setting pass rates on summative assessments; however, ultimate decisions about quality of performance should be based on the expertise of policy makers, practitioners, and assessment developers (see **Reliability**).

Appropriateness across Contexts

No significant correlations occurred among any context-related variables (i.e., grade level teachers taught, language diversity of students, or setting--rural, urban, suburban, inner-city).

The tasks piloted were focused mostly on 10th-, 11th-, and 12th-grade English classes. Although 64% of the teachers in Group I taught grades 6-8, there was no statistical difference between their scores and those of the high school teachers. In fact, in both groups, there was a slight negative correlation between tasks and teaching level, indicating that there was a negligible tendency for middle or junior high teachers to score higher than 9th-12th grade teachers. Teachers in both Group I and Group II did comment, however, that the tasks were less appropriate for them because the contexts specified were predominately high school.

Scorers recognized the bias towards high school contexts; however, they didn't believe that that should have a significant effect on teachers' scores. Most scorers noted that the tasks seemed fair in all ways, including in their representation of diverse student groups. One scorer, however, disagreed, responding:

> *Not too well-focused on teaching students who are ESL/LEP. It does probe a teacher's sensitivity, but in obvious ways...*
>
> *Another scorer perceived a bias in the types of literature represented in the prompts:*
>
> *The weaknesses of this assessment instrument included the lack of representative work of minorities and women in the tasks.*

Eighty-four percent of all teachers believed that the tasks were fair to English teachers of diverse student groups, including varying ability, levels, different ethnic groups, handicapped, and limited English speaking students. Nevertheless, some teachers listed specific weaknesses:

> *Not with regards to the handicapped.*
>
> ***Designing A Lesson Sequence*** *was an example of a culturally biased assignment. I would question the validity of asking students to write about receiving gifts from parents. I believe this is an experience that some of my students have never had.*
>
> *The assessments don't really address limited English students* ***OR*** *the different ethnic groups. Glosses over; almost stereotypes these students (Shame on you!)*
>
> *At-risk students need totally different formats, works of literature, matrix of interpersonal skills (from teacher), and a flexibility (on part of teacher), that other students don't need. Of course, this would apply to any teacher teaching any subject.*

> *Students of different ability levels pose different social and intellectual problems. The assessment didn't seem to account for regular/low academic students or gifted or special education students. Bilingual students were also left out.*

One policy question is raised regarding the various contexts that can be structured into simulation tasks. Once a teacher earns a secondary level credential, he or she is certified to teach any secondary level anywhere. Therefore, it may seem reasonable to expect that any teacher would be able to address student needs across grade level and regardless of the nature of the student population. However, being able to vary instruction based on shifting context is a skill known to be more characteristic of experienced teachers than beginning teachers (Leinhardt, 1983). The reality is that beginning teachers have relatively little experience across diverse settings. Can a second-year teacher, who may be competent in one setting, be expected to have student and curriculum knowledge relevant to other settings in which he or she hasn't taught?

Fairness across Groups of Teachers

There were no statistical correlations between scores on tasks and demographic variables--i.e., gender, age, ethnicity, undergraduate major, type of undergraduate institution, type of credential, number of English methods courses, years' experience, or bilingual training. There was a slight to moderate negative correlation between teachers' age and their scores on some tasks. No other meaningful patterns emerged.

One evaluation question to which teachers responded asked: Do you feel that this assessment is fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers? Eighty-eight percent of all teachers felt that the assessment was fair to teachers regardless of background. However, this statistic must be interpreted in light of the fact that only 14% of the teachers in the sample were minorities. One of the 56 teachers was Black, 5 were Hispanic, and 2 were Asian or Pacific Islander. Moreover, while there was no statistical correlation between ethnicity and scores across tasks, with such a small sample of minority teachers the differences would have had to have been dramatic for any to appear.

Of those who elaborated their responses regarding fairness across groups of teachers, the most common reference was to gender. Three Group I teachers expressed concern that only female teachers were portrayed in the tasks they completed. In most cases, gender of teachers was not divulged; however, it seemed clear to those teachers who mentioned it that there were no examples of male teachers.

Appropriateness as a Method of Assessment

Teachers were asked how the format of tasks they completed compared with other tests by which they've been evaluated (e.g., multiple-choice for CBEST and NTE, classroom observation) in terms of the tasks' "assessment ability."

Overwhelmingly, teachers preferred the simulation tasks to other pencil and paper tests, such as the CBEST and NTE. They felt that the kinds of information tapped by the structured simulation tasks seemed much more relevant to teaching. A number of teachers recognized the different purposes of the two tests. As one teacher wrote, "CBEST tests us as students; this test tests us as teachers."

A few teachers expressed concern about possible subjectivity in scoring the simulation tasks. These teachers felt that the simulation tasks tapped into important teaching skills, but they were afraid that reducing responses to "right" and "wrong" determinations would limit their value as assessment devices.

Many teachers commented that, while this was an improvement over other kinds of assessments they had taken, it did not substitute for classroom observation. A number of responses reiterated that the tasks ought to comprise a part of overall evaluation, with classroom observation included in assessment as well.

Whether these simulations are appropriate means of assessment depends on what they are intended to assess. Until validity of the tasks is determined by comparison between scores and observed teaching behavior, the simulations are best thought of as assessments of knowledge, not classroom performance. Some tasks have greater face-value potential as assessments of classroom performance than others. **Designing a Lesson Sequence** and **Responding to Student Writing** elicit responses that may realistically represent behaviors performed in actual classroom practice. Nevertheless, teachers commented that much more time was allowed to complete the simulation tasks than they actually expend in practice. Although the scenarios in **Responding to Typical Problem Situations** are realistic, responses are still hypothetical. **Stages of the Writing Process** and **Oral Presentation Skills** ask new teachers to critique strengths and weaknesses of other teachers' assignments, comments, and behaviors. Such evaluation has no correspondence to anything these teachers do in actual practice--new teachers don't formally evaluate other teachers. While completing simulation tasks appealed to teachers more than did multiple-choice tests, the appropriateness of these tasks is best evaluated in terms of the goals of the assessment, not in comparison with other assessments or forms of assessments.

Structured Simulation Tasks' Format

The structured simulation tasks are performance-based pencil and paper assessments. They therefore depend on clear directions and adequate time for completion. The following sections review the clarity of preparatory materials and task instructions, as well as the effectiveness of time allotted for each task.

Clarity of Preparatory Materials

Before teachers completed the structured simulation tasks, they were mailed a short (two-page) letter briefly describing the tasks and assessment procedures. Almost all teachers felt that the preparatory materials they received beforehand were adequate. Three teachers misconstrued the descriptions and expected that they would be witnessing actual lessons or oral presentations, rather than critiquing them on paper. Four teachers commented that examples of assessment questions would have helped, as well as suggestions about how to prepare.

Clarity of Task Instructions

A cover sheet for each task outlined the skills assessed, format of responses, context (e.g., "Twelfth grade English class in suburban high school...,"), time limit--with suggestions for pacing weight of scoring, and a checklist of enclosed materials. Teachers were asked to critique instructions for each task they completed. Scorers were also asked to critique the effectiveness of the tasks' directions based on how well teachers seemed to follow them. The following sections review the effectiveness of the instructions for each task, based on teachers' and scorers' observations.

Responding to Typical Problems. Only one teacher critiqued the directions for this task. This teacher commented that the directions give an example of how to respond to the scenarios using a bulleted format, but that the answer sheet is formatted in a way that suggests answers should be in paragraph form. Some teachers responded in bullets, others in paragraph form. This did not pose a problem for scorers.

Designing a Lesson Sequence. Ten of the twenty-eight teachers who completed this task expressed problems with the task's directions. Some comments were:

> *The directions didn't state whether the reading of the literature was to be included in our lesson sequence, or if we were to assume that the student had already read the selections. (Three other teachers expressed this same confusion.)*
>
> *Took too long to figure out what format they wanted [my response] in; I was afraid I did it all wrong, even after I did it!*
>
> *The only suggestion I have is that the directions [should] specify what skills have been previously covered. For example, the students were to complete the week with a compare/contrast essay. I didn't know whether an explanation in detail was necessary or whether this was a task the students already knew how to do.*

Although all teachers completed the lesson plans in an acceptable format, some teachers' low scores may be a reflection of the confusion expressed in the above comments. For example, respondents were rated down if their lessons focused on teaching a compare/contrast essay instead of teaching the themes of literature presented in the curriculum materials. This would have affected teachers who assumed students had no previous experience with a compare/contrast essay. Also, teachers felt that it wasn't clear whether students had read the literature assigned. If teachers assumed students had read the relevant literature, and thus didn't provide reading activities in their lessons, they would have been rated lower.

Responding to Student Writing. Although no teachers expressed difficulty with the directions of Part I of this task, four teachers expressed confusion about the directions in Part II. Teachers weren't sure whether they were supposed to write strengths and weaknesses about each teacher comment, whether they had to respond in the order of the lettered responses, or whether their responses for different comments could be combined. Two teachers weren't sure whether they could point out strengths and weaknesses in the same comment. Finally, one teacher would have liked to know "whether the student wrote the essay the way it was typed."

Developing Oral Presentation Skills. There were two parts for this task. Part I had teachers critique an assignment designed to foster students' oral presentation skills. In Part

II, teachers were presented with two 150-200 line transcripts of classroom sessions in which students presented oral reports. Teachers were to critique strengths and weaknesses in the hypothetical teacher's comments and behaviors in both scripts.

Apparently the task's directions were not clear to all teachers. One teacher wrote that she was in the middle of Part II before she realized that the script corresponded to the assignment elaborated in Part I. Another teacher noted that the directions were too complicated: "...I was looking at as many as four pieces of paper at one time." Another teacher emphasized that it needs to be made clear that teachers are to concentrate on the hypothetical teacher's involvement and not on student responses.

No evidence of teachers' confusion arose during the scoring. All teachers seemed to understand how to go about completing the tasks. One teacher did note in her evaluation that it would be best for teachers to skim the entire transcript before making comments, for better pacing. Since ten people did not finish commenting on the second transcript in Part II, this seems good advice.

Stages of the Writing Process. Only one teacher had problems understanding the directions for this task, not realizing that each of the three sub-tasks in Part II had a different answer sheet. Scorers were able to piece together the teacher's responses; however, the teacher lost points because answers that were intended to refer to two sub-tasks were only scored once.

Length of Tasks

Twelve of the twenty-eight teachers who completed **Developing Oral Presentation Skills** felt that the 80 minutes allotted was insufficient. Ten teachers did not finish commenting on the second transcript segment in Part II; others finished but commented that they felt rushed. Teachers suggested extra time allotments from 5 to 30 minutes, with most listing 15-20 extra minutes needed. However, one teacher suggested that instead of providing more time, provide less test, since the second script seemed redundant.

Two teachers suggested adding another 15-20 minutes for **Stages of the Writing Process**, which was allotted 60 minutes.

Three teachers in Group I thought that there was *too much* time allotted for both **Designing a Lesson Sequence** (allotted 120 minutes) and **Responding to Student Writing** (allotted 60 minutes). These teachers felt that, since they don't have that much time to plan lessons and respond to student papers in actual practice, allowing excessive time in a test situation compromised the tasks' validity. On the other hand, two teachers reported needing more time to complete **Designing a Lesson Sequence**, one of whom suggested an extra 30 minutes.

Cost Analysis

Administration and Scoring Cost Estimates

The Structured Simulation Tasks for Secondary English Teachers are administered in a large group setting using procedures common to standardized group administrations. The tasks can be administered by one or more persons with little or no training in the specific content of the assessment.

The largest component of the cost of this assessment is that of personnel. For this pilot test administration, a total of eight scorers were used to score six tasks, each of which was completed by 28 teachers. The scoring took place as follows: four scorers scored four tasks and part of one task in five days, and four scorers scored one task and part of another task in a day and a half. Although some of days spent scoring include time spent revising the scoring criteria and process, it does not include any time for training which we believe is necessary in order for the assessment to be fair and valid.

The amount of time required to score each task was more closely related to the number of its subparts than the length of time required by the teachers for its completion. The range of time it took to score each task was six hours to two days (this includes the time incurred for development work). Based on this experience, and assuming that future administrations would include some training, we estimate that approximately 2 1/2 days per scorer would be required to train and score roughly 30 responses to a single task. If a half-day assessment consisted of three tasks, it would take approximately 7 1/2 scorer-days to score 30 teacher assessments. According to this logic, four scorers should be able to score 100 teacher assessments resembling the tasks piloted in seven days, with periodic checks to insure that scorers are applying scoring criteria correctly. Assuming a cost of $150 per day for each scorer, this implies a cost of approximately $42 per teacher to train scorers and score an assessment. If these same scorers were used again for a similar task shell, the training time might be shortened, reducing marginally the total scoring costs.

Costs for test administration, duplication of materials, postage, travel, etc. would also need to be added to the costs for scoring the assessments. As we have outlined on other assessments, a cost of $30 per assessment for these activities assumes minimal travel costs for test administration. A summary of cost estimates for administering and scoring an assessment like this include:

Training and Scoring:	$42 per assessment
Administration/Other:	$30 per assessment
Total Administration and Scoring costs:	$72 per assessment

Development and Pilot Testing Costs

Table 6.6 shows costs for pilot testing by cost category which total $66,607. These data provide a rough indication of the magnitude of costs that would be incurred if a similar assessment were to be adapted for implementation.

Technical Quality

This section describes the process by which the assessment was developed, as well as the reliability and validity of the structured simulation tasks.

Development

The developer has submitted a final report, which summarizes development of the five structured simulation tasks (Klein, S. & Stecher, B, 1991). Developers were charged with creating a set of performance tasks that could be used on a licensing examination for secondary school English/Language Arts teachers. The tasks were conceived as discrete exercises, some or all of which would be used in conjunction with other types of assessments to make licensing decisions.

Two RAND researchers and six educators participated on the development team, which met between October 1990 and January 1991 (a seventh educator attended the first meeting only). Team members were recommended to developers at RAND by the state Department of Education. Most members had been active participants in other state-sponsored English/Language Arts assessment and/or curriculum projects. The development

TABLE 6.6

PILOT TEST COSTS FOR STRUCTURED SIMULATION TASKS FOR SECONDARY ENGLISH

Cost Categories	Pilot Testing
Staff-Salaries & Benefits	10,143
Consultants (Teachers, assessors, and other consultants)	37,127
Travel (Consultants and staff)	4,436
Other Direct Costs (Site rental, phone, duplication)	948
Total Direct Costs	$52,654
Indirect Costs	13,953
Total Costs	$66,607

team was headed by two RAND researchers. In addition, one team member was a university professor, two were district administrators, and three were classroom teachers. Five members of the team were women, three were men. One of the women was Black, all other team members were Anglo.

To facilitate future development of parallel tasks, the construction of each task revolved around a group of concepts, which developers call a "shell." Ideally, shells provide the following:

- a general description of the activity or types of activities that will be present in a task, (e.g., "grade a set of student papers that exhibit at least five of the following characteristics...") and the general directions to candidates;

- conditions that can be built into a task that candidates should attend to in specified ways and which can be scored with respect as to whether the candidate did or did not attend to them (e.g., recognition of good ideas in a poorly written essay);

- the types of materials candidates will receive (both in advance of the test and at the test site); and

- any special features of the context that ought to be explained.

In theory, many different case situations can be generated by the same shell. Developers can simply vary the characteristics of tasks--and therefore the specific knowledge or skills to be tapped--as well as vary the conditions in which those characteristics are embedded (i.e., student's ability or grade level). In practice, ideas outlined generally by the shell are continually modified as actual tasks are developed. The shell acts as a guide more than a blueprint. Full features of the shell become elaborated only after a task has evolved. In fact, at this stage, no shell for any of the five tasks has been fully developed. As a result, new tasks generated from the same shell may be only vaguely related.

Although no shell was fully developed, five tasks were; a sixth, **Conducting Student Discussion,** was left in an early stage of development due to resource limitations. The developed tasks were initially administered to eighteen prospective first- and second-year English teachers. Responses were scored and evaluated, which led to revision of scoring guides and, in some cases, the tasks themselves.

Reliability

Because the tasks were conceived of as discrete elements, Cronbach alpha coefficients were calculated for each task, as opposed to calculating an overall alpha coefficient for all three tasks teachers in each group completed. Three coefficient values are reported: one for raw scores, one for standardized scores, and one for adjusted scores, wherein raw scores were converted to pass/fail determinations based on scoring criteria and expert judgments (by scorers and/or the evaluator) of minimum acceptable competence. These values are reported in Table 6.7 (Note: an alpha coefficient could not be determined for **Designing a Lesson Sequence**, since there was only one score generated for each individual.)

Given the relatively early stage of development, there are some respectable alpha levels in some tasks: **Responding to Student Writing** Cronbach alpha = .678; **Responding to Typical Problems--Form B** Cronbach alpha = .690; and **Stages of the Writing Process** Cronbach alpha = .687 (all alpha statistics are for standardized scores). In fact, it would be surprising to find higher reliability coefficients for such open-ended tasks, especially at this stage of development. Reliability is, in part, a measure of internal consistency of scores. Developers note that in simulations such as those on the bar exam, fifty sample responses are used to calibrate raters. With twenty-eight responses per task and no prior administrations, it was impossible to calibrate raters before scoring actual responses.

In addition to consistency of scores, reliability is also a function of the number of sub-tasks scored. Generally speaking, the more scores to correlate, the higher the reliability. Thus, tasks such as **Stages of the Writing Process**, which generated ten scores, and **Responding to Typical Problems**, which generated six scores, would be expected to have higher reliability coefficients than **Oral Presentation Skills**, which generates three scores. Nevertheless, high reliability on these kinds of assessments is difficult to achieve, even in more advanced stages of development. The lower the reliability, the greater the chances of misidentifying candidates as being competent or incompetent.

Issues of reliability are inevitably tied to policy questions. Assessments developed to aid in licensure decisions necessarily define standards of minimum acceptable competence. Whatever the State considers necessary knowledge for teaching will need to be clearly defined and accurately evaluated. In any testing situation, the greater the focus on a facet of knowledge, the greater the chance the instrument will be reliable. Asking one question

TABLE 6.7

CRONBACH ALPHA COEFFICIENTS FOR THE SIX TASKS

Task	Coefficient Alpha		
	Raw Scores	Standardized Scores	Pass/Fail Scores*
Group I			
Responding to Typical Problems - Form A (6 scores)	.539	.543	.278
Responding to Student Writing (5 scores)	.616	.678	.618
Designing a Lesson Sequence**	-	-	-
Group II			
Responding to Typical Problems - Form B (6 scores)	.669	.690	.584
Stages of the Writing Process (10 scores)	.648	.687	.606
Developing Oral Presentation Skills (3 scores)	.235	.316	.401

*Pass/fail cut-offs were determined based on scoring criteria and scorers and/or the evaluator's expert judgements of minimum acceptable competenc

**Reliability coefficients depend on correlations of scores within subtasks; since only one score was generated for Designing a Lesson Sequence, a reliability coefficient could not be generated.

about a subject, for example, will almost always be less reliable than asking five or ten. In a letter to FWL (September 14, 1990), Dr. Stephen Klein, a developer of the Structured Simulation Tasks for Secondary English Teachers, clearly summarizes the challenges in developing a reliable assessment of minimum competence:

> *The issue of how much testing time will be allotted to licensing, how that time will be allocated across types of measures, how standards will be set, etc. are policy decisions that will most likely be driven by economics and politics rather than psychometric considerations. For instance, will California be willing to have a licensing program that fails more than 5% of the candidates? And, if 95% pass on their first try, should we demand the two to three days of testing time that almost certainly would be needed to insure that pass/fail decisions are made in a reasonably reliable fashion with respect to individual candidates?...The notion that a licensing test would consist of a few performance measures taken on one morning is way off the mark. Such a plan would not even come close to meeting professional standards set for test use.*

Consensus about what constitutes minimum acceptable competence could provide a necessary framework for further development of simulation tasks. Developing such a framework could be an extensive task, especially in the area of English instruction. Nevertheless, some successful efforts have been made, most notably by ETS in the process of developing the NTE, and at the Teacher Assessment Project at Stanford University. Determining the breadth and depth of standards to be assessed would enhance efforts to ensure reliable measurement of specified knowledge. Working from a specified framework would promote efficient and economical progress, should further development of these tasks be undertaken.

Inter-correlations Among Tasks

Correlations among tasks were calculated based on scores of the 28 teachers in each group. Results are reported in Table 6.8. Based on raw scores, two statistically significant relationships emerged ($p<.01$), although they are difficult to interpret. **Stages of the Writing Process** correlated with **Oral Presentation Skills** $r=.53$ ($p=.005$), and **Responding to Typical Problems** correlated $r=.51$ with **Responding to Student Writing** ($p=.006$). Converting to pass/fail scores based on scorer criteria and expert judgments, the latter correlation holds ($r=.57$, $p=.0015$), while the former doesn't ($r=.34$, $p=.240$).

TABLE 6.8

CORRELATIONS AMONG TASKS

Group I	1.	2.	3.
1. Responding to Typical Problems – Form A	1		
2. Responding to Student Writing	.51*	1	
3. Designing a Lesson Sequence**	.04	.24	1

Group II	1.	2.	3.
1. Responding to Typical Problems – Form B	1		
2. Stages of the Writing Process	.30	1	
3. Developing Oral Presentation Skills	.32	.53*	1

*p<.01

Validity

Typically, psychometricians speak of three kinds of validity: content, construct, and criterion validity. Content validity refers to the relevance of information assessed on a test to job performance. Teachers who complain about standardized tests such as NTE or CBEST often remark that "it has nothing to do with teaching." This is a comment about content validity. Content validity of the five simulation tasks in regards to teaching seemed high. Employing a development team whose members were familiar with classroom contexts helped ensure the relevance of skills and knowledge addressed. When teachers were asked their thoughts about the tasks overall, or in comparison to other assessments they had taken, their comments indicated the tasks have good content validity:

> *I feel this is a better assessment than the NTE or CBEST because it deals with reality.*
>
> *I enjoyed the assessment...The areas measured, particularly student writing, are pertinent to the position and allow me to exhibit positive skills.*
>
> *This assessment dealt more with the actual teaching situations than CBEST, which deals more with your own knowledge about various subjects.*
>
> *This tests skills I must have as a teacher.*
>
> *This assessment seems to be more in line with what a teacher actually does on the job. It seems to be an adequate simulation of life in the classroom.*

Construct validity refers to the extent to which a task measures what it says it measures. Some people criticize standardized math tests, for example, because completion of such tests relies heavily on a test-taker's reading ability. These tests, critics argue, measure reading ability as much as, or more than, math skill. Construct validity is difficult to define for simulation tasks based on job-related performance and not psychological constructs. However, one question that can be asked of each task is what, exactly, is it supposed to measure?

It was not always clear what facets of knowledge or skill were being scrutinized, especially when responses were scored holistically, based on a number of criteria. Is size or quality of a teacher's handwriting a legitimate facet of the *Responding to Student Writing* "construct"? Does **Responding to Typical Problems** measure of how teachers' actions 1) affect the attitudes of a particular student, parents, or other students, 2) reflect sound pedagogical principles, or 3) prevent backlash to the school as an institution--each of which were considered during the scoring of each scenario? Holistic assessment enabled scorers to weight one element more than another in order to arrive upon a single overall score. However, it will be difficult to establish defensible pass/fail assessments for tasks on a high-stakes test when the answer as to what it measures is "all of the above."

Criterion validity refers to how well performance on the test predicts actual job performance. Some developer materials refer to the simulation tasks as "performance tasks," and it is true that the simulations require teachers to "perform" during testing more actively than they might checking off true/false answers for three hours. However, policy makers should not confuse the concept of test-taking performance with classroom performance. Presumably, in that simulation tasks reflect the kinds of tasks teachers execute in practice, they have greater potential to predict actual performance. But this assumption is purely speculative.

The most promising tasks insofar as criterion validity are concerned are **Designing a Lesson Sequence** and Part I of **Responding to Student Writing**. But teachers themselves commented that the amount of time allotted to complete those tasks was unrealistically long, suggesting that actual performance differs from tested performance. Criterion validity of the other simulation tasks is less promising. Tasks that ask teachers what they might do if...(**Responding to Typical Problems**), and to critique other teachers' assignments, behaviors, and comments (**Oral Presentation Skills** and **Stages of the Writing Process**) have no direct correspondence to tasks beginning teachers execute in practice.

Performance on the simulation tasks *might* predict one's competence on the job, but that is only a hope. Until simulation results are correlated with other measures of classroom performance (which in themselves are difficult to collect reliably), they can only be thought of as assessments of teachers' knowledge about various facets of teaching.

On the other hand, one could see the simulations as assessments of what teachers know about teaching, rather than how they perform. Although there may be more economical ways to construct a knowledge test, some of the tasks seem to have the potential of being effective measures of teacher knowledge. **Responding to Typical Problems**,

Responding to Student Writing and **Designing a Lesson Sequence**, for instance, seem to have the potential of effectively eliciting essential understandings teachers have about various facets of teaching.

Summary

A development team headed by RAND researchers created five tasks to be considered as part of an assessment for secondary level English/Language Arts teachers. The tasks were: **Responding to Typical Problem Situations (Forms A and B)**; **Designing a Lesson Sequence**; **Responding to Student Writing**; **Stages of the Writing Process**; and **Developing Oral Presentation Skills**.

Fifty-six teachers, divided into two groups (n=28 in each group), piloted the developed materials. Each group completed three tasks (one of the tasks had two parallel forms). The tasks were evaluated as discrete exercises, rather than as a group of tasks comprising a single assessment. Furthermore, evaluation analyses assumed that the tasks were formulated to help make defensible summative (i.e., pass/fail) decisions for teacher licensure.

No significant correlations were found between task scores and any contextual or demographic variables. That is, teachers' scores did not seem to be influenced by gender, age, undergraduate major, type of institution where trained, number of English methods courses taken, type of credential held, or level and type of students taught. No effects were found for ethnicity; however, the sample of non-Anglo teachers was too small (8/56) to detect any but the most dramatic differences, if they had existed.

Ninety percent of the teachers felt that the preparatory materials and site arrangements were adequate. Some teachers expressed satisfaction that they didn't have to travel long distances to reach the test site. The need for a good amount of working space necessitated some teachers having to be re-located to a different room at one of the five test sites.

Teachers and scorers felt that the tasks were relevant to teaching, many expressing enthusiasm about the prospect of being assessed by these kinds of tasks in the future. A number of 6th-8th grade teachers felt that tasks were biased in favor of high school instruction. However, no statistical differences were found between scores of high school and junior high or middle school teachers.

Teachers and scorers agreed that the tasks were appropriate for teachers of all students, although some described exceptions, such as handicapped students, limited-English speakers, low-ability or gifted students. Most scorers and teachers believed the tasks to be appropriate for all teachers, regardless of years of experience or background. Teachers disagreed about how many years' experience would be necessary to perform adequately on the tasks, with suggestions ranging from 0 to 3 or more. One scorer felt the materials were not sufficiently representative of literature from diverse ethnic groups. Teachers and scorers commented that completing the tasks successfully would not depend on ethnic background; however, this perception must be interpreted in light of the fact that so few teachers represented non-Anglo backgrounds.

Directions on most tasks were clear. **Designing a Lesson Sequence** and **Developing Oral Presentation Skills** were mentioned most often as needing more simplified directions. Most teachers felt that there was sufficient time to complete each task, with the expectation of **Developing Oral Presentation Skills**, which 10 of the 28 teachers did not finish. Three (of 28) teachers commented that *too much* time was allotted for **Responding to Student Writing** and **Designing a Lesson Sequence**. These teachers noted that extended time limits in the test situation didn't reflect the minimal amounts of time given to complete these tasks in practice.

Taken together, the simulation tasks fairly thoroughly cover the *English/Language Arts Framework* and California Standards for Beginning Teachers. As would be expected, specific tasks vary in their coverage. Only two of the nine key facets of the *English/Language Arts Framework* are only partially addressed "*encouraging students to read widely*" and "*devising a diverse assessment program.*" Two elements of the framework, "*modelling effective use of all language arts skills*" and "*each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students*" were not directly addressed. Presentation Skills and Cognitive Outcomes of Teaching, two of the California Standards for Beginning Teachers, are not directly addressed. Task coverage of the *English/Language Arts Framework* and California Standards for Beginning Teachers will ultimately be influenced by the specific tasks candidates complete, how those tasks are developed, and the extent to which scoring guides are adhered to during scoring.

Scoring of the tasks revealed significant challenges for further development. Correlations among sub-tasks ranged from r=-.16 to .53, the average correlation being r=.16. Relatively low correlations would be expected among sub-tasks measuring different kinds of knowledge, or distinct facets of an underlying body of knowledge. Low correlations can also occur as a result of scoring inequities. Greater evidence exists that low correlations among sub-tasks are a reflection of scoring inequities more than they are accurate assessments of different facets of teacher knowledge.

Scoring methods varied across sub-tasks from analytic to holistic to a combination of the two. Scoring criteria were not always applied consistently, especially when holistic scoring was used. Sometimes inconsistency resulted from reasonable disagreements regarding acceptable practice. Other times, raters were not clear about the criteria being used in scoring. In all cases, scoring decisions were arrived upon by consensus, which was heavily influenced by interpersonal dynamics.

Reliability of the scores within sub-tasks ranged from Cronbach alpha =.316 for **Developing Oral Presentation Skills** to Cronbach alpha =.690 for **Responding to Typical Problems--Form B** (based on standardized scores). Reliability statistics are promising, considering the relatively early stage of development in which they were obtained. Not having a sample of papers with which to calibrate raters' decisions before actual tasks were scored would tend to lower reliability coefficients. Reliability also depends upon the number of scores entered into the equation. Tasks that generated more sub-task scores generally wound up with higher reliability coefficients.

Interpretation of teachers' scores on the tasks depends upon whether one sees scores as a reflection of teachers' knowledge or a reflection of the test's ability to assess it. Quality of responses--as judged by scorers--varied greatly from one task to another, and even from one sub-task to another. In order to make determinations about teachers' performance on tasks, decisions need to be made about the standards the assessment is meant to address, including defining the purposes of the assessment, what is designated as "minimum acceptable competence."

The tasks themselves have a high degree of content validity, which is to say that they are related to the kinds of knowledge important for teaching. It is not always clear, however, precisely what facet or facets of that knowledge is being assessed by a particular sub-task. Lack of limited and clear assessment objectives may have contributed to ambiguities that arose during scoring. Finally, no evidence exists that these tasks are in any way predictive of actual classroom performance. Insofar as the simulations can be seen as valid reflections of teachers' knowledge about teaching, some tasks seem particularly effective (**Responding to Typical Problems**, **Responding to Student Writing**, and **Designing a Lesson Sequence**).

Recommendations

The following recommendations are intended to guide future task development. The recommendations are based on analyses contained in this evaluation. Assumptions behind the analyses are that the goals of development are to create simulation tasks that can be used in conjunction with other means of assessments to make defensible summative (pass/fail) teacher licensing decisions.

Administration

- Should the structured simulation tasks become integral to licensure decisions, considerations will have to be made regarding the logistics of administering the assessment to large numbers of people on the same day, while providing the necessary surface space to ensuring that candidates are able to work independently.

Development

Scoring of these tasks presents great challenges to future development. Since problems in scoring are inextricably linked to development issues, recommendations related to both are presented below:

- Policy makers, in conjunction with qualified educators, can use the pilot test results from Group I and Group II to help determine standards, for the purposes of assessment, of "minimum acceptable competence." Some skills and abilities, though important, may not be assessable because of lack of consensus regarding what constitutes effective practice. Once it is determined what can and should be assessed, and once standards have been defined, they can become a framework for task development, including specified assessment objectives and scoring criteria.

- Careful consideration needs to be made regarding what scoring method--analytic, holistic, or a combination--is most appropriate for each task, given the goals of the task and the assessment overall. An analytic method has the advantage and disadvantage of weighing criteria equally. Holistic assessments have the advantage of being more valid intuitively, but possibly at a sacrifice to reliability, especially when consideration of a number of different criteria result

288

in a single holistic score. A combination method can be a check on either of the two methods, but can also carry with it both their disadvantages.

- In any system that relies on consensus to arrive upon a final score, a protocol should be developed and adhered to closely during scoring. The purpose of the protocol would be to mitigate the effects of interpersonal dynamics on the scoring process, especially in cases where scores are reached by consensus. The protocol should specify the responsibilities of a scoring team facilitator, whose role would be to 1) conduct training that ensures that all scorers fully understand all scoring criteria, and 2) monitor discussions to ensure all scorers are given an equal voice in scoring decisions. Ideally, the facilitator would be familiar with issues worked through during test development, and be available to facilitate scoring of different administrations. The facilitator ought to be sensitive to issues of interpersonal communication, including interaction styles predominant across different ethnic groups. Facilitators should not have a vote, except possibly in cases where agreement cannot otherwise be reached. The scoring protocol should specify when breaks should be taken, which ought to be at least once in the morning, during lunch, and at least once in the afternoon. The scoring observed by FWL staff was a draining process, and scorers' ability to focus weakened after long, uninterrupted stretches.

- Tasks should be constructed to yield multiple scores. The more scores a candidate has, the more statistically reliable an assessment is likely to be. Also, costs of administration (in terms of dollars spent per amount of information gained) go down when the ratio of scores generated to time used to get them is high. A task such as **Designing a Lesson Sequence**, which takes two hours to complete, is of limited value if it yields only one score, *regardless of how valid the task is or how effective the scoring*. In an excellent article that reviews many of these issues, Dr. Ed Haertel (1990) explains, "The *validity* of a test that chops up teaching into tiny segments may be suspect, but its *reliability* is likely to be quite high" (p.7). Of course, scoring costs rise if it takes more time to determine more scores, but costs may not rise proportionately to scores generated, and the value of the test increases dramatically.

- A serious effort must be made during development, piloting, and scoring, to incorporate the perspectives of diverse ethnic groups. In response to a similar set of tasks (see "Structured Simulation Tasks for Secondary Life/General

Science Teachers," Chapter 3), Dr. Sharon Nelson-Barber elaborated the need to recognize how notions of acceptable practice might vary according to ethnicity of students and teacher. Nelson-Barber noted the example of how an effective Black teacher's emphasis on strong adult leadership could differ from mainstream characterizations of good teaching, which tend to view the authority role as one of guiding and facilitating (Hollins, 1982; Delpit, 1988; Foster, 1989; Ladson-Billings, 1989).

Nelson-Barber also points out that certain culturally sanctioned teasing or "put-downs" built upon shared backgrounds have been very effectively used with Black inner-city college students (Foster, 1989). In **Oral Presentation Skills,** there is an example of a hypothetical teacher urging a student to speak up, saying (encouragingly) "Mike, you're a great big, Black boy. Now let's hear that great big voice." Labelling the comment as inappropriate earned candidates a point. However, the directions do not stipulate the ethnicity of the teacher. Conceivably, a Black teacher could have envisioned making the comment in ways that would have seemed appropriate. This is not to say that the example above was incorrectly scored. The point is that some teacher responses that could reflect sound practice in some circumstances and contexts might be considered inappropriate by development and scoring teams comprised of educators that represent only a limited range of ethnic backgrounds and teaching contexts.

Athanases (1991) addresses these issues in his discussion of the Teacher Assessment Project at Stanford University. "Effective teaching...includes culturally sensitive methods of working with diverse student groups...What works in one community of learners might not prove appropriate in another" (p.1). Assessments such as the structured simulation tasks institutionalize standards of acceptable practice. Only by significant representation of teachers from diverse ethnic backgrounds and teaching experiences can such professional standards be determined equitably.

- If any of the tasks are intended to be predictors of classroom performance, some validity studies must be conducted. The developers recommend a "known-group" validity study, where teachers with reputations for high performance complete the tasks. Such a study has merit, but should be combined with observation-based assessments as well.

- If any of the tasks are intended to be predictors of knowledge of teaching, the validity of scoring criteria must be explored, based on judgments of a wide range of experts. If scoring criteria are developed by a limited number of people on a development or scoring team, they will necessarily represent limited views of acceptable practice. If the assessment is going to have any value, it must allow for a range of responses based on valid though differing training and experience among professionals.

Conclusion

The Structured Simulation Tasks for Secondary English Teachers have the potential to assess teachers' knowledge and/or performance on a global level. Because the tasks allow teachers to combine knowledge and skills from a range of teaching domains (i.e., subject area expertise, pedagogy, knowledge of students, and knowledge of the institution), they provide a valuable understanding of ways teachers approach teaching complex tasks similar to those they manage each day. However, the effectiveness of any one approach depends heavily on contextual factors that are specific to the classroom, the school, and the broader parent and policy community, as well as responsiveness to anticipated or unanticipated outcomes of various decisions. Each standardized structured simulation task, e.g., lesson planning, focuses on some aspect of a beginning teacher's ability to teach a specific topic to a specific type of students, but the results of a single task are not generalizable within the type of task, topic or students. The assessment developer argues that with a large enough sample of tasks, topics, and type of students, a teacher's general teaching ability with respect to the subject area can be measured, though no diagnostic information would be available.

CHAPTER 7:

SECONDARY ENGLISH ASSESSMENT: ASSESSMENT CENTER ACTIVITIES

One of the more innovative assessment prototypes pilot tested, the Secondary English Assessment: Assessment Center Activities was developed at San Francisco State University, San Francisco. The prototype can truly be described as an alternative performance-based assessment as it includes four very different activities, each of which requires the teacher to demonstrate (or "perform") a different skill or ability.

Although all of the activities are designed to assess a teacher's approach to language and literature learning in a multicultural, multilingual context, the first three activities are conducted during a half day at an assessment center, and the fourth takes place over a three-month period in the teacher's classroom. The assessment center activities ask the teacher to demonstrate performance abilities in reading, writing, speaking, listening, responding to literature, evaluating student writing, and explaining language concepts.

- **Activity A, Responding to Student Writing** -- (approximately 50 minutes) This activity focuses on the teacher's skill in responding to student writing in a particular context. The teacher is given two samples of student writing and, for each sample, is asked to 1) write directly on the student sample addressing the student writer, and 2) analyze the student's text on a separate evaluation form, writing for a peer audience. Upon completion of the activity, which can be administered by a proctor, the teacher's responses are evaluated by at least one assessor.

- **Activity B, "Fishbowl" Discussion of Literary Work** -- (approximately 50 minutes) In this group exercise, the teacher is asked to demonstrate his/her skills of literary interpretation and collaborative learning. To prepare for the activity, the teacher reads a designated short story, responds to the story in an informal log, and prepares questions for discussion. Then, at the assessment center, the teacher participates in an oral discussion of the story with three other new teachers who have read and prepared the same story. The activity is simultaneously administered and scored by two assessors.

- **Activity C, Speaking of Language** -- (approximately 50 minutes) This activity focuses on a teacher's skill in impromptu oral performance. Sitting on a panel with three other new teachers, the teacher is asked to give an impromptu oral presentation (approximately five minutes long) on a topic pertaining to language and literature in a multicultural society. The teacher makes the presentation in response to a question based on a set of readings provided for the assessment. After the presentation, the teacher answers one follow-up question posed by a fellow panelist. The activity is simultaneously administered and scored by two assessors.

The fourth activity of the Secondary English Assessment is the Classroom Portfolio which is prepared by the teacher in his/her classroom. A description of this activity is as follows:

- **Activity D, Classroom Portfolio** -- (to be completed during a three-month period) This activity evaluates a teacher's skills in three areas: planning and implementing a teaching unit, responding to student work, and reflecting upon his/her experience in teaching the unit to gain insight for further teaching. The teacher plans and conducts a three- to six-week teaching unit in which the classroom activities are unified by a single focus. To document the teaching activities, the teacher compiles a classroom portfolio which consists of various interrelated components (e.g., weekly log, materials and assignments given to students, samples of student work with teacher responses). The completed portfolio is submitted to at least one assessor for evaluation.

For each of the four activities, the teacher's performance is evaluated at three levels: (1) according to specific criteria listed under a particular skill or category, (2) at the skill or category level, and (3) at an overall level. Figure 7.1, for example, is the response form used by assessors to evaluate a teacher's performance on **Activity A, Responding to Student Writing**. As depicted on the form, a teacher's performance is evaluated according to specific criteria corresponding to two skills or categories: (1) Response Strategies, and (2) Analysis of Writer and Text. The teacher is given a rating for each of the criteria in both categories, a rating for each skill or category, and an overall rating. As is also indicated on the form, at

FIGURE 7.1

Response Form A
RESPONDING TO STUDENT WRITING

Candidate's Name: ______________________________ Date __/__/__

Evaluator: ______________________________

Part I. *In comments addressed to student writer, the Candidate:*

	Very Strong				*Very Weak*	*Evaluator Comments*
A. Conveys reader's interest by responding to writer's meaning (content, ideas, information) not merely evaluating technique	4	3	2	1		
B. Demonstrates understanding of writer's goals and purposes	4	3	2	1		
C. Responds in a way that would promote confidence in future writing attempts	4	3	2	1		
D. Provides accurate and useful feedback on technical aspects of the writing	4	3	2	1	NA	
E. Makes helpful suggestions for revision or future writing	4	3	2	1	NA	
OVERALL RESPONSE STRATEGIES	4	3	2	1		

Part II. *In analyzing the student text for colleagues, the Candidate:*

A. Adequately describes writer's purpose/goals and method (identifies intended genre, task definition)	4	3	2	1	
B. Identifies effective features of the text::					
a. content	4	3	2	1	NA
b. structure (focus, organization)	4	3	2	1	NA
c. development	4	3	2	1	NA
d. style/voice (syntax, diction)	4	3	2	1	NA
e. mechanics	4	3	2	1	NA
C. Identifies problematic features of the text:					
a. content	4	3	2	1	NA
b. structure (focus, organization)	4	3	2	1	NA
c. development	4	3	2	1	NA
d. style/voice (syntax, diction)	4	3	2	1	NA
e. mechanics	4	3	2	1	NA
D. Draws sound inferences about general strengths of writer	4	3	2	1	
E. Makes sound suggestions about what writer would benefit from learning	4	3	2	1	
OVERALL ANALYSIS OF WRITER AND TEXT *(Additional comments on reverse)*	4	3	2	1	
OVERALL RATING	4	3	2	1	

Key to Rating Scale:
4 = *Definite Strengths in this area;* 3= *Some strength in this area;* 2 = *Lacks strength in this area;* 1 = *Serious weaknesses in this area;* NA = *not applicable in this instance.*

each level the teacher's performance is rated along a four-point scale, with a rating of "4" indicating a very strong performance and a rating of "1" being a very weak performance. All of the ratings are made in a holistic manner and are not interdependent. That is, the category ratings are not a composite or sum of the criteria ratings, and the overall rating is not a composite or sum of the category ratings.

After completing the four evaluations, the original assessment design required that a summative evaluation be made called a Competency Profile. The Competency Profile was designed to serve as a synthesis of the four evaluations and was to include a recommendation regarding credentialing. Due to a variety of reasons (e.g., time constraints, design of scoring response forms), the Competency Profile was not piloted as part of the Secondary English Assessment.

The first three activities (i.e., **Activities A, B, and C**) were pilot tested during the spring and summer of 1990, and are discussed in this chapter. The portfolio activity, which allows approximately three months for a teacher to plan and teach a unit, and compile the portfolio, was administered in the fall and winter of 1990, and is discussed in Chapter 8.

The administration, the content, and the format of the first three activities of the Secondary English Assessment are discussed below. The content and format sections of the report contain information from the teacher and assessor evaluation forms, as well as information and analysis of scoring results. Following these three sections are sections on cost analysis and technical quality. The chapter concludes with an overall summary together with recommendations for further steps in exploring the feasibility and utility of assessment activities such as these in California teacher assessment.

Administration of Assessment Center Activities

Beginning with an overview of the administration of the three assessment center activities of the Secondary English Assessment, this section provides information on the following: logistics (e.g., identifying the teacher sample, scheduling the activities), security, assessors and their training, scoring, and perceptions of the assessment activities by teachers, assessors, and FWL staff.

Overview

The three assessment center activities of the Secondary English Assessment were administered on August 11, 13, and 14, 1990 from approximately 9:00 a.m. to 1:00 p.m. each day. Four trained assessors and a FWL staff person administered the activities in two conference rooms at a hotel in San Francisco. Although 20 teachers were scheduled to participate, one teacher had to cancel at the last minute due to personal problems.

As shown in Table 7.1, of the 19 participating English teachers, the majority were Caucasian (non-Hispanic) females teaching at the high school level. An equal number (8) of teachers came from schools in northern and southern California, and three teachers came from schools in the central valley. Approximately two-thirds of the teachers were participating in the CNTP-sponsored teacher support projects. A little over one third of the teachers were teaching in inner city schools.

Logistics

Administration of the three assessment center activities entailed numerous logistical activities. First, there were activities for all the assessments such as identifying a teacher sample, recruiting and training assessors (who also served as scorers), scheduling and making arrangements for the teachers to be assessed, sending orientation materials to the teachers, and acquiring evaluation feedback from the teachers and the assessors. In addition to these activities, there were two other important logistical activities relevant to this assessment: recruiting trainers and developing the training for assessors, and extensive revision of the developer's original orientation materials.

Identifying the teacher sample. As mentioned earlier, Table 7.1 presents information about the teacher sample for this assessment. As was the case with other assessments, it was necessary to recruit Non-Project teachers in addition to Project teachers in order to have a sample that represented different regions of the state. It was also hoped that by recruiting Non-Project teachers, the teacher sample might have a better representation of different ethnic groups; however, only one of the Non-Project teachers was identified as non-Caucasian. The Non-Project teachers, almost all of whom identified themselves as suburban and urban teachers, were recruited by calling school districts and asking for names of first-year and second-year English teachers. All teachers, Project and Non-Project, were offered $300 to participate in the three assessment center activities and to complete a portfolio.

TABLE 7.1

PILOT TEST PARTICIPANTS
SECONDARY ENGLISH ASSESSMENT
(Number of Teachers = 19)

Location	No. of Teachers		Teacher Characteristics
	Project	Non-Project	
Northern California	4	4	16 Caucasian, non-Hispanic; 2 Hispanic; 1 Asian or Pacific Islander
Southern California	6	2	6 Male; 13 Female 12 High School; 7 Junior High
Fresno	3	-	6 Suburban; 5 Urban (not inner city); 7 Inner City; 1 N/A
Total Number of Teachers	13	6	

Recruiting and training of assessors. Four experienced high school English teachers were recruited and trained to serve as assessors and scorers for the Secondary English Assessment. Two of the teachers had previously administered and scored the activities during an August, 1989 workshop conducted by the developers. The other two teachers had no previous experience with the assessment.

The four teachers attended a two-day training session given at FWL. The training was given by two trainers, both of whom were recruited by FWL staff from the pool of experienced English teachers who had previously administered the assessment in 1989. The trainers worked with a member of FWL staff in designing the training. (For more information, see the section, "Assessors and Their Training.")

Scheduling/Arranging the assessments. The original 20 teachers were scheduled over three days. Because two of the assessment activities are administered to groups of four, the number of teachers scheduled for each day had to be a multiple of four. The final schedule was as follows: 8 teachers on Saturday, August 11; 8 teachers on Monday, August 13; and 4 teachers on Tuesday, August 14. Whenever possible, the teachers were given a choice of the day for assessment.

In addition to scheduling the teachers, air and/or hotel arrangements were made for some of the teachers so that they could attend the San Francisco administration site. Arrangements were also made to reimburse teachers for assessment-related expenses (e.g., parking, air shuttle).

Developing and sending the orientation materials. The developer provided the state with a package of orientation materials which were used in the developer's 1989 pilot test (i.e., the August workshop). These materials, however, were deemed by CTC and SDE staff, as well as by FWL staff, as in need of extensive revision for the 1990 pilot test. A member of FWL staff, with the assistance of CTC and SDE staff members, revised all of the orientation materials (as well as the assessment booklets used on assessment day) for Activities A through D.

The revised orientation materials comprised an orientation handbook sent to all of the teachers before the administration of the assessment. The materials described each assessment activity (including the portfolio activity), the criteria by which the teachers would be evaluated, and any preparatory activities they needed to complete before assessment day (e.g., keeping a log, completing a practice assignment). The handbook also included all the reading materials needed for each assessment activity (e.g., a short story, a set of articles).

Collecting evaluation feedback. FWL staff designed two evaluation feedback forms on which the teachers and the assessors could give their thoughts and opinions about the assessment. The teachers filled out the evaluation form immediately after they completed the three assessment activities. The assessors completed their evaluation forms on the last day of the assessment.

Security

For the developers of the Secondary English Assessment, test security for the assessment center activities during the pilot test primarily meant one thing: The group of teachers on whom the prototype was pilot tested had equal naivete about the test form. This security was achieved by the fact that not one of the teachers who participated in the pilot test indicated that they were familiar with the assessment in any way.

In the future, if some version of the prototype were adopted by the state for licensure purposes, the developers suggest that "it would be essential to involve many people in an introductory training session to ensure equal understanding of and preparation for this alternative form of assessment." Furthermore, the developers advocate "a full scale dissemination effort to guarantee all teacher training programs equal access to models of the prototype, for integration into teacher preparation as befits the individual campuses." (Both quotes taken from a March 6, 1990 letter written to the CNTP co-directors from one of the developers of the assessment.)

Even if these measures are followed, however, the developers recognize that not all teacher candidates will receive the same quality training to prepare them for the assessment. It is for this reason that the developers designed the assessment so that the inequality of preparation could be offset by the provision of extensive materials and instructions for individual self-preparation.

Other security measures to be considered if the prototype is adopted would be the collection and storage of the following assessment materials: (1) the assessment booklets for each activity, (2) the assessment response forms (i.e., the scoring sheets for each activity), and (3) any preparatory materials (e.g., reading logs) for the activities completed by the teachers. These materials would need to be retained for a minimum number of years, enough to cover the period in which teachers could appeal decisions, or to meet statutory requirements.

Assessors and Their Training

As mentioned earlier, four assessors were trained to administer the Secondary English Assessment: Assessment Center Activities. This section describes some characteristics of the assessors, describes the training, and presents the perceptions of the training by the assessors and FWL staff.

Characteristics of the assessors. The four assessors trained to administer this assessment were all experienced high school English teachers who also had experience in formal writing assessment programs involving holistic scoring of writing samples (e.g., Bay Area Writing Project) and other language arts organizations (e.g., CLP, CATE). Two of the assessors had previous experience with the Secondary English Assessment and two did not. The two assessors who participated as assessors in the August, 1989 workshop conducted by the assessment developers were both males, one Caucasian, the other Asian. The two novice assessors were Caucasian females who had been recommended by one of the assessment trainers. (When administering the assessment activities, the four assessors worked in teams of two, each team consisting of one male and one female, and one experienced and one inexperienced assessor.) All four assessors were from northern California.

Training. Although the August 1989 workshop conducted by the assessment developers included training of assessors, that training was limited as it did not have the benefit of videotapes of the activities, performance data on the teachers, and scoring data from the assessors. The August workshop supplied the above components, and, as a result, new training procedures and activities were developed.

The two trainers for the assessment met with a FWL staff person for one day at the end of June to discuss the design of the new training. In addition to being asked to design the training so that it incorporated videotapes, performance data, and scoring data, the trainers were informed of the changes made to the orientation and assessment materials (i.e., assessment booklets), and that those changes would also need to be incorporated into the training. The trainers then met together during the summer and designed the training. They met again with the FWL staff person before the scheduled training to review their design package.

The training was conducted at FWL in San Francisco on two days: Thursday, August 9, and Friday, August 10, 1990. The first day of training was devoted to **Activity A, Responding to Student Writing**. Because the administration of this activity did not require a trained assessor (i.e., the directions were self-administered), all four assessors were only trained to score this activity, not administer it. Training began with the assessors completing Activity A as if they were a candidate and then sharing and discussing their responses. Training continued with a discussion of the scoring criteria as listed on the scoring response form, and then practice in scoring teacher responses. Assessors scored three pairs of teachers' responses; after each pair they shared their ratings and discussed any differences and/or difficulties they may have had.

On the second day of training, the assessors were divided into two groups: two assessors were trained by one trainer to administer and score **Activity B, "Fishbowl" Discussion of Literary Work**, and two were trained by the other trainer to administer and score **Activity C, Speaking of Language**. Training for each group began with a review of the activity and the activity's scoring criteria. The rest of the training for the activity consisted of watching videotapes of teachers participating in the activity, and then scoring the teachers' performances. After each round of videotapes and scoring, assessors tallied their responses, discussed differences, and came to a consensus. The training ended with the assessors reconvening as a group and discussing the overall assessment procedures for the three days of administration.

Perceptions of training. When asked if the training they received was "very good," "adequate," or "insufficient," all four assessors responded, "very good." One assessor commented, "I felt the length, content, and format of the training were perfect!" Another assessor remarked, "Format and content seem exactly right."

All four assessors described two aspects of the training as being the most useful: (1) practice in scoring sample responses, and (2) discussion of the scores and evaluation criteria. Other aspects of the training that assessors found useful were the use of videotapes (one assessor described them as providing "excitement, interest, and reality"), the practice in doing **Activity A, Responding to Student Writing**, and the "establishment of the feeling that each assessor's opinion was valued."

All four assessors also had suggestions for improving the training. The suggestions given and the number of assessors who gave them were as follows:

Provide scoring guides that "would nail down the distinctions for each of the score points" (i.e., 4, 3, 2, and 1)
(2 assessors)

Provide more specific guidelines in the format and detail of writing comments on the scoring response sheets
(1 assessor)

Extend the training by one day or half a day
(1 assessor)

Based on our own observations of the training and on the performance data from the activities (which is discussed in the "Assessment Content" section), FWL staff agree that the training could be improved by following the above suggestions. In particular, the training should be revised to include an assessor's handbook which describes the scoring process in detail and specifically provides concrete examples whenever possible of (1) the distinctions between score points, and (2) the way in which comments are to be written on the scoring response sheets. By extending the training by a half day or more, both of these components could be addressed more thoroughly.

Scoring

The scoring system for the Secondary English Assessment: Center Activities is a holistic process which relies heavily on the assessor's professional judgement. Although teachers are rated along specific evaluation criteria described for each activity, these criteria serve solely as guides to help the assessor arrive at a holistic judgment for each skill or category and for the overall rating.

The scoring process for this administration was conducted as follows. For **Activity A, Responding to Student Writing**, each teacher's responses were scored independently by a pair of assessors. The responses were scored in the afternoon, after the teachers had completed all of the assessment activities. For **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language**, the teacher's responses were scored during and immediately after the activity. Each teacher was independently scored by the pair of assessors who administered the activity.

After every activity, as time allowed, the pairs of assessors discussed their scores to determine the degree of consensus. When differences were notable, they discussed their reasons for their scores and tried to achieve consensus; they did not, however, change their original scores.

In the afternoon, after every teacher had been scored on all three activities, FWL staff reviewed the comments made by the assessors on the scoring response forms and noted which comments were inappropriate or insufficient. Assessors whose comments were too subjective, for example, were instructed to write future comments as objectively as possible. Assessors who gave lower ratings (i.e., a "2" or "1") but did not provide comments which explained the ratings were asked to provide such comments in the future.

As the scoring procedures of this assessment are part of the assessment format, more information about the scoring procedures and the assessor's response to these procedures is provided in the section of this chapter titled, "Assessment Format."

Teacher, Assessor, and FWL Staff Perceptions of Administration

All 19 of the participating teachers expressed satisfaction with the arrangements (e.g., scheduling, room arrangements) made for the administration of this assessment. Comments about the arrangements ranged from "Excellent!" and "Great!" to "Very easy." One teacher who found the arrangements to be reasonable did add, however, that "unfamiliarity with the city created nervousness, as did the need to get up early and process information."

Although all four assessors also found the arrangements to be satisfactory, there were also some suggestions for improvement. One assessor suggested that the number of teachers assessed in **Activity C, Speaking of Language** could be increased from four to six. Another assessor wrote:

> *A list of items to remember to do, step by step, for the assessor would make it more consistent. Even with my own notes, I forgot to hand out questions once!*

And still another suggestion was to tape record the oral presentations made in **Activity C** to help answer questions about performance.

In response to the above assessors' comments, it is our assumption that the first assessor's suggestion means that the two assessors who administer Activity C could just as easily score six teachers as four. Based on the scoring data for the activity which includes comments made by the assessors on the scoring response forms, FWL staff hesitate to agree with this suggestion. It is our belief that the scoring process for this assessment needs revision and it is unclear whether these revisions would make it easier or more difficult to score teachers' performances.

As for the second suggestion, this information could be provided for every activity and should be included in the Assessor Handbook which we recommend be developed for the assessment. The tape recording of oral presentations could also be considered, especially to provide more examples during future training to represent the different ratings.

The teachers and assessors were also asked to comment on the amount of time allotted for the administration of the assessment activities. Approximately two thirds of the teachers thought the time allotted for each of the activities was sufficient, and one third did not. All four assessors though the amount of time allotted for the activity which they administered (i.e., **Activity B or C**) was sufficient. Since the timing of the activities is also a feature of the assessment format, this issue will be discussed more completely in the section, "Assessment Format."

Assessment Content

The content of the three assessment center activities of the Secondary English Assessment focuses on a teacher's skills in the following areas:

a) responding to and analyzing a student's writing;

b) literary interpretation and group collaboration; and

c) oral performance with regard to issues of language and literacy in a multicultural society.

The three areas were deliberately chosen by the assessment developers to represent aspects of competence which are not now assessed or are under-assessed during the credentialing of English teachers. The three areas focus primarily on competence in content pedagogy and subject matter knowledge, as opposed to general pedagogical competence.

General pedagogical competence is assessed, however, by the fourth activity, the Classroom Portfolio, described in Chapter 8.

In addition to its focus, another important aspect of the assessment content is its context. As was mentioned earlier, the content for each of the three activities incorporates a multicultural, multilingual context. For example, in **Activity A, Responding to Student Writing**, the teacher is asked to (1) read ten samples of student writing from a tenth-grade, multi-ethnic English class for context, and (2) respond to two samples of student writing, one of which is written in non-Standard English. For **Activity B, "Fishbowl" Discussion of a Literary Work,** the teacher is asked to read and discuss a short story written by an African-American author. **Activity C, Speaking of Language**, provides the teacher with a set of articles, taken from a variety of publications, about literature and literacy in the multicultural classroom to be read in preparation for giving an impromptu oral presentation on a related issue.

In the following pages, the content of the Secondary English Assessment: Center Activities is discussed along the following dimensions:

- Congruence with the California English/Language Arts Framework and Handbooks;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the assessment activities;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, diverse student groups);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

Congruence with the California English/Language Arts Framework and Handbooks

FWL staff reviewed the Secondary English Assessment to see in what ways the three assessment center activities are congruent with California's *English-Language Arts Framework*, 1987. Because two of the assessment activities focus specifically on writing and literature respectively, we also looked at congruence of the assessment activities with California's *Handbook for Planning an Effective Writing Program*, 1986, and *Handbook for Planning an Effective Literature Program*, 1988.

Table 7.2 describes the ways in which the different activities are congruent with the framework and handbooks. As is evident from the descriptions, all of the activities are congruent in some way with the framework and handbooks, but none of the activities are strongly congruent. **Activity A, Responding to Student Writing**, for example, addresses only one stage in the development of a student's composition skills--i.e., the teacher's evaluation of the student's writing. It does not assess a teacher's skill in providing various writing opportunities, in helping students write for various audiences, in teaching students how to write for a purpose, to revise and edit their writing, etc.--all of which are discussed in the *Handbook for Planning an Effective Writing Program*. One way in which the activity could be made more congruent is if the activity were expanded to include questions which address how the teacher might, in addition to evaluating the student's first writing, lead the student through the various stages of writing.

Similarly, **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language**, address only one component of an effective English-Language Arts program--i.e., modeling by the teacher of important English-Language Arts skills. Unfortunately, there is no guarantee that because a teacher's literary analysis or oral performance skills are strong, s/he can teach those skills to students. To make these activities more congruent would require substantial revision of the activities. **Activity B**, for example, might be revised so that the teacher views videotapes of students discussing a short story (which the teacher has read and then responded to in a reading log) and then is asked to critique the discussion along the dimensions of literary interpretation and group process. (If such a major revision were not deemed acceptable, then at the very least, the activity could be revised to include a question as to what oral language activities the teacher could conduct to help students better understand the meaning of the short story.) **Activity C** might be revised in a similar way: the teacher views a videotape of a student giving a speech and then is asked to comment on the student's content, organization, and delivery (which are the same criteria by which the teacher is currently evaluated when s/he gives the speech). Both of these revisions would result in more congruency with the state framework and handbooks because both would put a greater focus on a teacher's skill in responding to students' abilities versus demonstrating their skill in activities that have little or only an indirect relationship to teaching students.

BEST COPY AVAILABLE

TABLE 7.2

CONGRUENCE OF THE SECONDARY ENGLISH ASSESSMENT WITH THE ENGLISH-LANGUAGE ARTS FRAMEWORK AND HANDBOOKS

Framework and Handbooks	Activity A, Responding to Student Writing	Activity B, "Fishbowl" Discussion of Literary Work	Activity C, Speaking of Language
English-Language Arts Framework, 1987, K-12	Addresses development of a student's composition skills.	Integrates all elements of language (listening, speaking, reading, and writing). Addresses the issue of modeling. Teacher models good listening, valuing of ideas, and encouragement of questions.	Integrates all elements of language (listening, speaking, reading, and writing). Addresses the issue of modeling. Teacher who can speak well encourages students to use words well and to speak effectively.
Handbook for Planning an Effective Literary Program, 1988, K-12	Addresses development of a student's composition skills.	Addresses the issue of modeling (reading, writing, and listening skills). Focus on central issues, interpretation of symbols, discussion of meaning, and argument of interpretation.	Fosters awareness of society. Activity C articles address issue of literature in a multi-cultural society. Models oral language skills.
Handbook for Planning an Effective Writing, Program, 1986, K-12	The topic selected for the students' first writing samples could be considered appropriately motivating for a first draft.	Incorporates the use of reading logos.	Incorporates the use of reading logos.

Extent of Coverage of California Standards for Beginning Teachers

The three assessment center activities of the Secondary English Assessment were examined by FWL staff to see how well they covered the California Beginning Teacher Standards which define levels of pedagogical competence and performance that California teacher credential candidates are expected to attain (i.e., Standards 22 to 32). The standards are reprinted below (in italics), along with an analysis of how the assessment activities correspond to each standard.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* This standard is addressed in a small way by **Activity A, Responding to Student Writing** which assesses a teacher's skill in responding to student writing "in a way that promotes confidence in future writing attempts." This standard is not addressed by **Activities B** and **C**.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* None of the assessment center activities addresses this standard. (It is, however, addressed by the Portfolio activity).

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities, and materials that are appropriate for students with diverse needs, interests and learning styles.* In **Activity A, Responding to Student Writing,** the teacher is asked to respond to writing samples from two different students, both of whom have different needs. The standard is not directly addressed by **Activity B** or **C**; however, **Activity C, Speaking of Language** indirectly addresses the standard through some of its questions which ask the teacher to explain his/her view about teaching literature in a multi-cultural classroom (e.g., teach "the Classics" or teach multi-cultural literature?).

Standard 25: Student Motivation, Involvement, and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* The motivation aspect of this standard is somewhat addressed by **Activity A, Responding to Student Writing** which asks the teacher to respond to student writing in "a way that would promote confidence in future writing attempts." This standard is not addressed by **Activity B** or **C**.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* In **Activity A, Responding to Student Writing**, the teacher is assessed on how well s/he communicates to students (via written language) about their writing. **Activities B** and **C** do not address this standard.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* In **Activity A, Responding to Student Writing**, the teacher is asked to evaluate samples of student writing. **Activities B** and **C** do not address this standard.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* This activity is not addressed by **Activity A**. **Activities B** and **C** also do not directly address this standard, but it could be inferred that a teacher would be unable to improve the ability of students in a class to evaluate information, think analytically, and reach sound conclusions, unless the teacher him/herself could do so--as is required by Activities B and C.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* This standard is addressed by **Activity A, Responding to Student Writing**, which asks the teacher to respond to student writing "in a way that promotes confidence in future writing attempts" (i.e., the student would want to continue to write and would feel good about him/herself as a writer). This standard is not addressed by **Activity B** or **C**.

Standard 30: Capacity to Teach Cross-Culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socio-economic differences.* This standard is not directly addressed by **Activities A, B**, or **C**. Indirectly, however, each of these activities touches upon this standard: **Activity A** requires the teacher to respond to a sample of student writing which is written in a non-standard dialect; **Activity B** requires the teacher to discuss a short story about African-American children written by an African-American author; and **Activity C** sometimes requires (depending on the question selected) the teacher to discuss issues about teaching literature in multi-voiced, multi-cultural classrooms. A teacher's capacity to teach cross-culturally could be inferred from his/her response to the student writing sample, the short story, and the literature issue, but this is not directly addressed by the current scoring criteria.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* This standard focuses on a teacher's ability to teach classes which span the range covered by the credential (i.e., grades K-8 or 7-12) or students at two or more ability levels (such as remedial and college preparatory classes). None of the activities are designed to assess this ability. This standard also addresses a teacher's ability to fulfill typical responsibilities of teachers such as meeting school deadlines and keeping student records, none of which are assessed by any of the activities.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* This standard is not directly addressed by any of the activities, although **Activity B, "Fishbowl Discussion of a Literary Work**, does assess a teacher's ability to communicate and cooperate with others in the discussion of a short story.

The extent of coverage by the Secondary English Assessment: Center Activities of the California Beginning Teacher Standards is summarized in Table 7.3. The table lists the assessment center activities which address each standard, and also describes the extent of coverage provided.

Job-relatedness

The 19 teachers who participated in the three assessment center activities of the Secondary English Assessment were asked if the skill areas chosen for the activities (i.e., (a) responding to and analyzing a student's writing; (b) literary interpretation and group collaboration; and (c) oral performance with regard to problematic issues in English education) are relevant to their job of teaching. All of the teachers but one responded positively, some offering comments such as the following:

> *Resoundingly so!*
>
> *In fact, more relevant than most districts are willing to commit time to.*

TABLE 7.3

EXTENT OF COVERAGE BY THE SECONDARY ENGLISH ASSESSMENT OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Assessment Center Activity Addressing Standards	Extent of Coverage
22: Student Rapport and Classroom Environment	-Activity A, Responding to Student Writing	Limited
23: Curricular and Instructional Planning Skills	-None	None
24: Diverse and Appropriate Teaching	-Activity A, Responding to Student Writing	Limited
25: Student Motivation, Involvement and Conduct	-Activity A, Responding to Student Writing	Limited
26: Presentation Skills	-Activity A, Responding to Student Writing	Limited
27: Student Diagnosis, Achievement and Evaluation	-Activity A, Responding to Student Writing	Limited
28: Cognitive Outcomes of Teaching	-None	None
29: Affective Outcomes of Teaching	-Activity A, Responding to Student Writing	Limited
30: Capacity to Teach Crossculturally	-Indirectly Addressed by Activity A, B, and C	Limited
31: Readiness for Diverse Responsibilities	-None	None
32: Professional Obligations	-Activity B, "Fishbowl" Discussion of Literary Work	Limited

A few teachers, however, qualified their "yes" answers by stating or suggesting that **Activity C, Speaking of Language** was not relevant. Commented one teacher,

> *Except for Activity C. Interesting content and issues; however, how can this impromptu [speech] on issues assess oral performance in the classroom? Delivering such information and to colleagues (not students) is different than speaking to students.*

Another teacher remarked,

> *I don't exactly see where Activity C fits into our daily activities, beyond the political question of being able to "defend" or lobby particular education theories.*

This latter remark was somewhat echoed by the one teacher who responded negatively to the question of job relevancy. This teacher found **Activity C** to be irrelevant because, "I don't deal with problematic issues except those I can control in my classroom."

Of the four assessors, all found the three assessment center activities to be relevant to a teacher's job. One assessor emphatically wrote:

> *Decidedly relevant! I would be very hesitant to hire a teacher who lacked the skills to do at least passably well on all of these activities. The abilities to analyze and comment on student writing, discuss literature, and to explain one's pedagogical philosophy and practice are crucial to teachers.*

Three of the assessors, however, judged **Activity A, Responding to Student Writing** to be the "most relevant" because, in the words of one assessor, it is "closer to actual classroom performance." Elaborating on the merits of **Activity A**, an assessor commented,

> *[Activity A] assesses one of the most important and pervasive activities of English teachers, and the two essays are as diverse as possible: black/white, female/male, extremely/barely literate, and diverse in style. Responses to both give a good picture of a candidate's knowledge of both pedagogy and subject matter in the teaching of writing.*

FWL staff agree with most of the above comments: the three assessment activities do seem relevant to a teacher's job, **Activity C, Speaking of Language** does seem the less relevant, and **Activity A, Responding to Student Writing** the most. However, with regard to **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language**, we also tend to agree with a teacher who acknowledged the indirect job relevancy of the activities, but who added,

> *They in no way would assess our actual job of teaching a class.*

In other words, although **Activities B and C** are indirectly relevant to an English teacher's job, and, it could be argued, part of a teacher's job, they do not in any way assess how a teacher responds to and teaches students, which is the teacher's job.

Appropriateness for Beginning Teachers

The appropriateness of the three assessment center activities of the Secondary English Assessment are discussed in this section from two perspectives: (1) the perceptions of the participating teachers and assessors, and (2) the teachers' performance on the assessment.

Perceptions. When asked if they had sufficient opportunity to acquire the knowledge and skills relevant to the activities in which they participated, the teachers' responses were mixed: 63% (12 of 19) said "yes," 21% (4 of 19) said "no," and 16% (3 of 19) said "yes and no." Several of the teachers who marked "yes," however, qualified their answer. These teachers said that they had sufficient opportunity to acquire the knowledge and skills relevant to the activities, but only because of their experience in the classroom. For example, one teacher wrote,

> *Yes, but mainly because I've had one year's experience already. Without that I don't think I would have done as well.*

Another teacher elaborated further, commenting both on her experience in the classroom and her lack of training at the university:

> *Personally, I had no preparation for the assessment of my skills in responding to student writing during my university course work in teacher training. Now, having taught for two years, I've developed my philosophy and approach with help from*

> *teachers I respect. But straight out of teacher training, this [activity] would have been assessing something I hadn't been taught.*

Some of the teachers who responded to the question with both a "yes" and "no" answer, also cited lack of training or exposure to the content of the activities as the reason for their dual answer. In explanation of the "no" part of her answer, one teacher commented on her lack of exposure to one part of the content of **Activity C, Speaking of Language:**

> *The only reason I marked "no" is because I personally have not been exposed to the current debate: canon vs. multicultural literature.*

Similarly, another teacher explained her dual answer as follows:

> *No, because my department, school and district spend almost no time dealing with these kinds of activities. Yes, because of my own interests.*

Finally, the majority of the teachers who responded "no" to the question also referred to their lack of experience in the classroom and lack of training in the skills areas assessed by the activities. Remarked one teacher,

> *As a new teacher, I think I need more experience in these three areas to be considered "skilled." A college degree doesn't necessary give me the knowledge to teach nor does a teacher preparation program. I think a lot of knowledge and skill comes from experience.*

Two other teachers with "no" answers zeroed in on their lack of experience/training in a particular assessment area:

> *I need practice in the process of collaborative thinking and cooperative groups [Activity B]. I have a general understanding, but lack experience in which to draw information from. I was taught to be an authority which is inconsistent with collaborative learning.*

> *I feel my responses to student writing [Activity A] and my ability to be clear about issues in articles [Activity C] need lots of work.*

The four assessors were also asked if they thought that a beginning English teacher would have had an opportunity to acquire the knowledge and skills needed to respond to each activity in an adequate manner. None of the assessors gave an unqualified "yes"; three qualified their responses, and one gave a dual answer much like that of some of the teachers. The major qualification linking all of the assessors' answers was that of the quality of the teacher's preparation program. One assessor remarked,

> *They should have had opportunity to develop these skills and knowledge--if they haven't, I think their deficiencies are an indictment of teacher training programs. Teachers should be trained to respond to student work, analyze literature, and read and synthesize research.*

Said another assessor more succinctly,

> *Yes, if they have gone through a good teacher training program. No, if they haven't.*

Offered another assessor,

> *Some teacher training programs obviously don't require the knowledge and skills necessary, but that's the fault of the program, not the assessment. Maybe one use of [the assessment] is to evaluate these programs.*

In conclusion, while a slight majority of the teachers believe they have had the opportunity to acquire the skills and knowledge measured by the assessment activities, many of the teachers and all of the assessors are not so sure. The dissenting teachers and those expressing uncertainty explained their answers by citing lack of experience in the classroom and/or a lack of training in the activities skill areas. The assessors justified their answers by focusing on the possibility of a teacher not having had a good teacher preparation program (i.e., one that trained the teacher to respond to student work, analyze literature, and read and synthesize research).

Performance on assessment. FWL staff analyzed the teachers' performance on each of the three assessment activities to see if the beginning teachers participating in this assessment had acquired the knowledge and skills measured by these activities. Specifically, FWL staff looked at the teachers' overall ratings for each activity, as well as the corresponding skill or category ratings. Because each teacher was rated by two assessors, the ratings from both assessors were included in the analysis. Although the rating scale included four possible ratings, ranging from a high of "4" to a low of "1", the ratings were not designed with pass/fail characteristics. For our purposes, however, we interpreted the "4" and "3" ratings (4 = definite strengths in this area; 3 = some strengths in this area) as "pass" ratings, and the "2" and "1" ratings (2 = lacks strength in this area; 1 = serious weaknesses in this area) as "fail" ratings.

Table 7.4 shows the number of teachers receiving each rating in the evaluation categories (including "overall") for each activity. In the first activity, **Activity A, Responding to Student Writing**, approximately 63% (12 of 19) of the teachers clearly passed (i.e., they received an overall rating of "3" or "4" from two assessors), and 16% (3 of 19) clearly did not pass (i.e., they received an overall rating of "2" from two assessors). The remaining 21% (4 of 19) of teachers were given a "2" rating by one assessor, and a "3" rating by another. None of the teachers received an overall rating of "1."

In the second activity, **Activity B, "Fishbowl" Discussion of Literary Work**, 89% (17 of 19) of the teachers clearly passed, and none of the teachers clearly failed. Two teachers received mixed ratings: one teacher received a "3" and a "2"; another teacher received a "4" from one assessor and no rating from the other assessor. None of the teachers received an overall rating of "1."

In **Activity C, Speaking of Language**, 68% (13 of 19) of the teachers clearly passed, 16% (3 of 19) clearly did not pass, and 16% received a mixed rating (i.e., a "3" and a "2"). None of the teachers received an overall rating of "1."

Overall, then, at least half of the teachers clearly passed each activity, with the greatest number of teachers passing **Activity B, "Fishbowl" Discussion of Literary Work**. Moreover, approximately 79% (15 of 19) of the teachers clearly passed at least two of the three activities, with 42% (8 of 19) of the teachers passing all three activities.

TABLE 7.4

THE NUMBER OF TEACHERS RECEIVING EACH RATING IN THE EVALUATION CATEGORIES FOR EACH ACTIVITY

ACTIVITY A, RESPONDING TO STUDENT WRITING

	Clearly Passed				Clearly Did Not Pass	
Evaluation Categories	"4"	"3"	"3 & 4"	"3" & "2"	"2"	Other
Response Strategies	1	6	4	6	2	
Analysis of Writer to Text	0	6	3	3	3	4 (missing a rating)
Overall	1	10	1	4	3	

ACTIVITY B, "FISHBOWL" DISCUSSION OF LITERARY WORK

	Clearly Passed				Clearly Did Not Pass	
Evaluation Categories	"4"	"3"	"3 & 4"	"3" & "2"	"2"	Other
Interpretative Process	9	5	1	1	1	2 ("2" & "4")
Group Process	5	8	5	1	0	
Overall	8	7	2	1	0	1 (missing rating)

ACTIVITY C, SPEAKING OF LANGUAGE

	Clearly Passed				Clearly Did Not Pass	
Evaluation Categories	"4"	"3"	"3 & 4"	"3" & "2"	"2"	Other
Content	6	4	3	3	3	
Plan	3	7	3	2	3	1 (missing rating)
Delivery	5	7	5	2	0	
Overall	4	6	3	3	3	

The performance data discussed above seems to support the teachers' perceptions that they have had the opportunity to acquire the knowledge and skills measured by the assessment activities, at least for **Activity A, Responding to Student writing** and **Activity C, Speaking of Language**. Approximately 63% of the teachers indicated that they have had the opportunity to acquire the necessary skills and knowledge to pass the activities, and 63% and 68% of the teachers respectively clearly passed **Activities A and C**. The teachers did much better on **Activity B, "Fishbowl" Discussion of Literary Work** with over three-fourths of the teachers clearly passing. This marked increase in performance is not surprising, however, since some form of literary analysis is usually taught beginning in the 7th or 8th grade (if not before), and talking about books with others is also often an integral part of the curriculum beginning about then.

The performance data also support the teachers' responses to the question of which of the assessment activities, if any, could be passed after student teaching and before teaching a classroom of their own. Approximately 89% of the teachers (17 of 19) named **Activity B, "Fishbowl" Discussion of Literary Work**. **Activity A, Responding to Student Writing** and **Activity C, Speaking of Language** were named by less than 50% of the teachers (8 and 7 respectively). Similarly, when asked which of the activities, if any, could only be passed by teachers with more than two years experience in the classroom, five teachers each (or 26%) named **Activity A** and **Activity C**, while only two teachers (10%) named **Activity B**.

In conclusion, the analysis of the teachers' performances on the three activities seems to suggest that **Activity B, "Fishbowl" Discussion of Literary Work** may be more appropriate for beginning teachers than **Activity A, Responding to Student Writing** and **Activity C, Speaking of Language** primarily because beginning teachers seem to have had greater opportunity to acquire the knowledge and skills measured by **Activity B**.

Appropriateness across Contexts

In order to determine if the teachers and assessors believe the three assessment center activities of the Secondary English Assessment are appropriate for teachers across contexts, we specifically asked them to comment on the assessment's appropriateness for teachers of diverse students groups (e.g., different student ability levels, different ethnic groups, handicapped or limited-English students, different school/community settings). Approximately 89% (17 of 19) of the teachers responded positively to the question; one teacher disagreed, and one teacher was undecided. Of the teachers who responded positively, one teacher affirmed,

> *They are skills that are needed no matter the particular circumstance of the teacher.*

The teacher with the dissenting opinion remarked that "only Activity A was appropriate," because the other two activities put the teacher in the position of the student (e.g., discussing literature in a group) rather than teaching the student.

The following sections discuss the issue in more detail with respect to teachers of different grade levels and of diverse student groups (e.g., ethnic diversity).

Grade level. In this assessment pilot test, none of the teachers or assessors made any reference to the inappropriateness of the assessment for teachers at different grade levels.

Analysis of the rating results, however, indicate that there may be some differences among teachers of different grade levels according to the activity. For example, of the three teachers who clearly did not pass (i.e., received a "2" rating from two assessors) **Activity A, Responding to Student Writing**, all taught at the junior high/middle school level. One other middle school teacher received a "3" and a "2" rating for this activity. Thus, of the seven junior high/middle school teachers participating in **Activity A**, more than half did not clearly pass.

In **Activity B, "Fishbowl" Discussion of Literary Work**, no teacher clearly did not pass, and one high school teacher received a "3" and a "2" rating. In **Activity C, Speaking of Language**, three teachers clearly did not pass, two of whom were middle school teachers. In addition, three teachers received a "3" and a "2" rating, two of whom were middle school teachers.

Thus, the performance data on the teachers seems to suggest that junior high/middle school teachers may be less well prepared than senior high teachers for **Activity A** and **Activity C**. (It is hard to imagine why this is so, however, unless, in general, the more skilled secondary teachers gravitate toward and are hired at the high school level.)

Diverse students. As mentioned above, a clear majority of the teachers believe the assessment is appropriate for teachers of diverse student groups. One teacher added, however,

> *The classics would be impossible to teach ESL students. The story could be told, but the reading could not actually happen.*

The teachers' belief that the assessment is appropriate for teachers of diverse student groups is strengthened by the fact that they are all teachers of diverse student groups. For example, all of the teachers who participated in the assessment taught in classrooms where at least some students spoke a language besides English. In addition, more than 50% (11 of 19) of the teachers taught in classrooms where four or more languages were spoken.

The assessors were also asked to address the issue of student diversity, but in a slightly different way. The assessors were asked to comment how the assessment activities address a beginning English teacher's ability to work with diverse students. All of the assessors agreed that **Activity A, Responding to Student Writing** does a good job of addressing this issue. Commented one assessor,

> *Activity A definitely addresses the candidate's ability to effectively communicate with students of very different backgrounds and writing abilities.*

The assessors were more mixed in their comments about **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language**. Three of the assessors tended to agree with the following remark made by one assessor about **Activity B**:

> *Activity B can reveal something about the teacher's abilities in this area if the story used for the activity is one that requires cross-cultural knowledge to be understood, as is the case with "The Lesson."*

The assessors agreed that "The Lesson," the story used in this pilot test, does a good job of raising issues relevant to cultural diversity. One assessor, however, disagreed with the other three, stating that **Activity B** is not a suitable way of assessing a teacher's ability to work with diverse students.

Activity C, Speaking of Language was also perceived by three of the assessors as being able to "give some indication of a teacher's ability to work with diverse students" because of the focus of the readings and the topic choices for the oral presentations.

Activity C was still considered less suitable than **Activity A, Responding to Student Writing** however, because, in the words of the following assessor,

> *Activity C indicates theoretical grounding in the issues, but Activity A shows the actual practical reaction to "non-standard" papers.*

Our consultant on cultural diversity, Sharon Nelson-Barber brought up another perspective regarding the assessment's appropriateness for teachers of diverse student groups. Barber examined some of the teachers' responses to **Activity A** and was struck by one teacher's analytical comment about the student's text which in effect said that the student had the potential of "straying away from the topic" in future writings. Barber stated that this comment "suggests a lack of knowledge about/experience with students who employ different rhetorical strategies." That is, research shows that there are considerable organizational differences between white children's oral narratives and those of certain racial/ethnic groups. Black students, for example, tend to be more episodic and white students more "topic-centered" in their oral narratives (Michaels S. and Cook-Gumperz, 1979). Native American students have also been documented as sometimes exhibiting different organizational patterns in their narratives (Colley, R. and Lujan, P., 1982).

As the assessors did not respond to the above teacher's comment, we do not know if they know the information presented by Barber and chose not to react to the teacher's comment, or if they are unfamiliar with the information. An assessor familiar with the research might rate a teacher lower than an assessor unfamiliar with the research. Thus, in order to be a fair assessment for teachers of diverse student groups, it seems that the assessors would have to be familiar with the current research on students of different racial/ethnic groups as it pertains to the topic of the assessment activity (e.g., student writing).

Fairness across Groups of Teachers

When asked if they felt the assessment is fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers, the majority of teachers and assessors said "yes." Explained one teacher,

> *No matter your background, you should be able to do these things.*

Only two teachers responded negatively, one of whom commented,

> *I know of a young man who is Hmong and studying to be a teacher. I think Activity C might be difficult for him. I think these different cultures need to be represented.*

Another teacher, who chose not to answer the question with a "yes" or "no," agreed with the above teacher, remarking that the "oral language area might prove difficult to ESL speakers."

One of the assessors also perceived the assessment to be biased towards fluent English speakers, but stated,

> *Fluency in English is required, but that doesn't strike me as unreasonable.*

Finally, another factor of the assessment was pointed out by an assessor who noted that "obviously a candidate whose hearing or sight was impaired would require an adapted assessment."

Aside from these somewhat obvious factors, our consultant on cultural diversity, Sharon Nelson-Barber, brought up two other issues. First, she points out that, in **Activity A**, the teachers are directed not to mark every error in the student's writing "unless that is your practice on such occasions." She acknowledges that the intent of this direction might be "to discourage the teachers from feeling compelled to mark errors simply to demonstrate they see them," but she wonders if a teacher who elects to frequently mark errors might receive a lower rating than a teacher who does not. In fact, she cites the following assessor's comment on one of the teacher's response forms as an example of probable bias against teachers who focus on mechanics: "Response is heavily focused on mechanics. Likely to undermine the warmth of the response."

In response to this assessor, Barber reminds us of the following:

> *Many black teachers view the teaching of skills as essential to their students' survival--that moving ahead to mastery of mainstream language means practice with skills. Thus, grammar, punctuation, spelling, etc. are precisely the features some black teachers are likely to highlight in the assessment.*

The other issue Barber addresses pertains to **Activity B, "Fishbowl" Discussion of Literary Work.** In this activity, teachers are expected to discuss a short story in a small group format. Barber reminds us that, "as was discerned during the Stanford Teacher Assessment Project's (TAP) small group discussion exercises, not all group members participate equally in group discussions, even when they know their patterns of interaction will be noted and rated." It is therefore important that assessors are aware of differences in language use across groups, as well as non-verbal communicative cues and interactive styles (e.g., some participants take the role of leaders, others of followers). Such knowledge is especially important when the group members are of different cultures, explains Barber, for "it is all too easy for a person to feel 'something is wrong' in interactions with people of a different background without really knowing what is causing this feeling."

Thus, once again it seems that the fairness of this assessment for teachers of different groups could depend heavily on the knowledge of the assessors.

Our analysis of the scoring results as they pertain to different groups of teachers indicates that females tended to receive higher overall ratings than the males for **Activity A, Responding to Student Writing** and **Activity B, "Fishbowl" Discussion of Literary Work,** and lower overall ratings for **Activity C, Speaking of Language.** In **Activity C,** for example, all three teachers who clearly did not pass were females (it should be noted, however, that twice as many females as males participated in the pilot test).

Further analysis of the scoring results reveals that teachers who described themselves as teaching in suburban locations tended to receive higher ratings for all of the evaluation categories for all three activities. Our very small sample of minority teachers (3) tended to receive higher ratings than the non-minority teachers on **Activity B, "Fishbowl" Discussion of Literary Work,** lower ratings on **Activity A, Responding to Student Writing,** and a mixture of lower and higher ratings for **Activity C, Speaking of Language.** (For more information on trends of differences between teachers with different characteristics, see the section, "Technical Quality.")

Appropriateness as a Method of Assessment

In addition to evaluating the appropriateness of the Secondary English Assessment: Center Activities for beginning teachers, and its appropriateness across contexts and groups of teachers, the teachers and assessors were asked to evaluate the appropriateness of the method of assessment, and to compare it with other methods of assessment which they have experienced.

Appropriateness. The teachers were asked if they thought the three assessment center activities are an appropriate way of assessing (1) general teaching skills, and (2) skills in teaching English classes. Approximately 58% (11 of 19) of the teachers responded positively to the first question, and 74% (14 of 19) to the second. Many of those teachers who did not think the assessment center activities are an appropriate way to assess general teaching skills defended their answers by saying that the activities did not require any teaching. Some comments illustrative of their viewpoint are as follows:

> *You can't assess someone's teaching skills unless you see them teach. I feel like a failure in this, yet in the classroom I am developing confidence, and have been told that I am a good teacher by my peers who have seen me teach.*
>
> *I think the written evaluation is good for English assessment, but teaching is not only knowing your content area, discussing a book with friends, or talking in front of a polite group. It is interaction with students. A teacher needs to know how to guide a discussion, keep students focused, deal with the interruptions, and alter plans when things fall apart. I think knowing the philosophies advocated in the articles helps, but training for war is not the same as being in the middle of an ambush.*

This point of view was also shared by some of the teachers who did not think the activities were an appropriate way of assessing skills in teaching English classes either:

> *We were not asked to TEACH anything...Our interpretation of the "The Lesson" may reveal how well we can read and analyze literature, but our discussion doesn't necessarily prove we can teach interpretation skills.*
>
> *The second exercise [Activity B, "Fishbowl" Discussion] placed us in a student's role, not a teaching role (as if we were back in college). The third exercise [Activity C, Speaking of Language] is much like a speech contest. No teaching is involved...not at all actual lesson related.*

Even one of the teachers who responded positively to both questions, added a qualification, saying,

> *Except in the case of Activity C, Speaking of Language. What did you intend to assess? Speaking skills? Teaching delivery? I'm not sure it adequately assesses my ability to deliver a lesson to a student.*

Finally, from a teacher who answered the questions with a "yes" and a "no":

> *Best way to judge someone's teaching effectiveness and the ability to teach students is to actually see it. Many educators have all the correct terms and such but choke in front of the room.*

The assessors were also asked to comment on the appropriateness of the assessment as a way to assess general teaching skills and skills in teaching English classes. All four stated that the assessment activities were an appropriate way of assessing skills in teaching English classes. Like the teachers, however, there was more uncertainty as to whether the activities were an appropriate way of assessing general teaching skills. Only two of the assessors responded with a definite "yes"; one gave a qualified "yes," and the other commented as follows:

> *I'm not completely sure that a person might not teach well despite everything else...unless his performance on all three [activities] were abysmal.*

In conclusion, while a majority of teachers and all of the assessors think the three assessment center activities are an appropriate way of assessing skills in teaching English classes, many of the teachers and some of the assessors reject the notion that these activities are an appropriate way of assessing general teaching skills. In particular, the teachers faulted the assessment center activities for not assessing a teacher actually teaching anything.

Comparison of activities with other assessments. All of the teachers were asked to compare the three assessment center activities of the Secondary English Assessment with other assessments with which they have been evaluated (e.g., multiple-choice exams such as CBEST and NTE Specialty Areas Tests, classroom observations during student teaching).

Of the 18 teachers who responded to the question, 72% (13 of 18) commented favorably about the assessment, many stating that the assessment is better than the NTE and/or CBEST tests. Some of the teachers gave high praise to the assessment because, unlike other assessments, it provided an opportunity to learn:

> *This has been the most delightful and helpful of the assessments I've gone through. This method is a learning opportunity in itself (in dialogue with other new teachers, etc.). I think the CBEST is absurd as a realistic method of assessing a teacher's capabilities.*
>
> *I think this assessment has far more value than any of the above-mentioned techniques [because] we are all learning and developing as we are being assessed.*

Other teachers praised the assessment activities because they were performance-related:

> *Are you kidding? You couldn't even compare an assessment with this much individual attention and hands-on performance with traditional pen and paper tests.*
>
> *CBEST is useless. This assessment judges who I am more--it's more personal. Can I speak well? Can I write well? How do I sound in front of a group? Can I communicate? CBEST assesses none of this.*

A few teachers, while not commenting negatively about the assessment or making unfavorable comparisons, still championed classroom observations as the best way to assess teaching competency for the obvious reason that they assess a teacher actually teaching:

> *Classroom observation and student feedback are the two most realistic ways of assess teaching competency...this assessment is all related to teaching...but I don't know if it assesses a person's teaching skills.*

> *I feel the best way to evaluate a teacher is in front of the classroom so you can actually see the reactions and responses to the students.*

Finally, one teacher best summed up the comparison of the assessment activities with CBEST, NTE and classroom observations as follows:

> *No comparison, CBEST and NTE are like playing the trivial pursuits literature game. I think that classroom observation during student teaching is extremely important, but doesn't always reflect the teacher's thinking about teaching.*

Assessment Format

The format of the Secondary English Assessment: Center Activities has a dual nature: there is the format of the assessment as a whole (i.e., the assessment as a single entity), and there are the distinctly different formats corresponding to each activity. In this section, the formats of the three assessment center activities are sometimes discussed separately and sometimes together, depending on the focus of the analysis. In discussing the assessment's preparation materials, for example, the assessment is primarily discussed as a single entity. When discussing the clarity of the assessment's rating forms, however, each of the three activities is looked at separately. The format section is primarily based on the comments of the teachers and assessors, as well as the perceptions of FWL staff.

Clarity of the Teachers' Preparation Materials

In preparation for the three assessment center activities of the Secondary English Assessment, each teacher received an Orientation Handbook sent in advance of the assessment. This handbook included a section on each assessment activity, with each section including the following elements:

- overview
- evaluation criteria
- preparation activities
- sample instructions
- evaluation response form
- preparation materials

When asked how thoroughly they read the handbook, 89% (17 of 19) of the teachers said they read it carefully. One teacher admitted to skimming it, and another read some parts carefully and skimmed others. All but one of the teachers said that the assessment activities, the aspects of teaching being evaluated, and the preparation activities were described clearly in the handbook. The one dissenting teacher thought the assessment activities and the preparation activities were described clearly, but that the aspects of teaching being evaluated were not.

The majority of teachers (12 of 19) were also satisfied with the information presented in the handbook. Teachers praised the handbook for being "very complete," "clearly organized," and "easy to read." One teacher remarked,

> *The format was excellent and left little room for misunderstanding.*

In addition to praise for the handbook, there were suggestions for improvement. These suggestions ranged from "use a metal spiral binding instead of a plastic one" to "do not include all the teacher evaluation forms--they detract from the purpose."
Several teachers commented that the section on **Activity C, Speaking of Language** needed improvement. One teacher thought the instructions for the activity were repeated; another teacher felt the instructions were not explicit enough (i.e., this teacher thought the directions should specify that the teacher should prepare ahead of time an outline for a response to all of the questions which may be asked at the assessment center). Still another teacher wrote,

> *Activity C was confusing because there was so much to read and interpret. I found myself less clear about what the issues are at the end of the reading.*

The above teacher's comment about **Activity C** is an important one, and is discussed further below. As for the other suggested improvements, FWL staff do not agree that the teacher evaluation forms (i.e., the response forms on which the assessors rate the teacher's performance) should be eliminated. The inclusion of these forms provides the teachers with the exact criteria by which their performance is judged on each activity. FWL staff acknowledge, however, that because of the way the handbook was organized, some of the instructions for the activities were repeated, and it is possible this redundancy could be eliminated or reduced.

After being asked their opinions about the handbook, the teachers were asked if they had any difficulties with the preparatory work described in the handbook. Approximately 50% (9 of 19) of the teachers answered "yes," the majority of them (6 of 9) citing difficulty with the preparatory work required for **Activity C**. Specifically, the difficulties described by the teachers all related to the set of five articles which they were required to read in preparation for their impromptu presentation at the assessment center. These articles were described by some of the teachers as "dry," "difficult to read," and "of little interest." One teacher remarked,

> *Some articles were jargon-laden and a little difficult. I had to reread parts to understand them clearly.*

Other teachers commented on a specific article with which they had difficulty. In all, three of the five articles were singled out by at least one teacher as being dry or difficult to understand.

(There were also teachers, however, who felt just the opposite about the articles, as evidenced by the following comment:

> *I thought the articles in the reading (Roemer, Hirsch, et al.) were terrific--they made me think, get inspired. I plan to Xerox Roemer's article for my colleagues at school.)*

The other difficulties experienced by teachers related to the reading log required in preparation for **Activity B, "Fishbowl" Discussion of Literary Work**. The three teachers who experienced difficulties with this activity all expressed uncertainty as to what they should write in the log. Said one teacher,

> *I was a little confused about what we were to do for Activity B. I wasn't really sure what was expected of me in my reading log.*

Another teacher described a lack of experience in keeping a reading log, and thus had trouble writing the minimum amount (i.e., one typewritten page).

After commenting on the handbook and the difficulties they experienced with the preparatory activities, the teachers were asked if they had any other comments about the preparatory work required for the assessment. While two teachers cited a shortage of time to complete all the work, and other teachers repeated some of the problems described above, a little over one-fourth of the teachers commented favorably about the preparatory process, most of them referring indirectly to the preparation work required for **Activity C, Speaking of Language**. A sample of these comments follows:

> *It was important to have the preparation book because it allowed time for processing. If I had to evaluate and assess an article on the spot, my success would be lower than when I had proper time to prepare, think, and process information.*
>
> *The preparatory work was excellent to prepare me for the assignment.*
>
> *It's a good idea to give preparatory work because research is a new experience and the preparation work prepared me. It was like reading a play before going to see it in the theater.*

This last teacher's comment is particularly noteworthy as it offers a possible explanation as to why some of the teachers had a difficult time reading some of the articles for **Activity C**. Reading research articles does not seem to be a common activity of beginning teachers--nor probably of teachers in general. Thus, the language used in the articles may be intimidating to some teachers or at the very least be unfamiliar. In fact, this set of articles was (1) compiled by several high school English teachers, and (2) selected over another set of articles for inclusion in the handbook because these articles were deemed to be more readable and interesting! While FWL staff acknowledges that some of the articles (e.g., Hirsch's article) are written in what could be called educational research, we believe the content of these articles is very important and beneficial for teachers to read.

In summation, while the majority of teachers read the Orientation Handbook carefully and were satisfied with the information presented, several teachers suggested improving the section on **Activity C, Speaking of Language**. In addition, almost one third of the teachers expressed having difficulty with the preparatory work required for **Activity C**. The difficulty they cited most often was that all or some of the articles were difficult to understand or not very interesting--a difficulty which may be the result of a lack of experience in reading research articles. Several teachers also expressed difficulty with preparing the reading log for **Activity B, "Fishbowl" Discussion of Literary Work**. These teachers expressed confusion as to what they should include in their logs. Finally, despite some teachers' difficulties with the preparatory work required for the assessment, there were other teachers who appreciated and praised the work for preparing them for the assessment.

Appropriateness of Time Allotted for Each Activity

Each of the three assessment center activities was allotted approximately 50 minutes. Each activity, however, utilized this time in a very different way. **Activity A, Responding to Student Writing**, is broken into two parts. In the first part, the teacher reads and responds to the two student writing samples. In the second part, the teacher evaluates/analyzes the two writing samples on separate forms. For this activity, it was suggested that the teacher allot approximately 10 minutes per essay in part one, and 10 minutes per essay in part two. (Obviously, this gives the teacher an extra 10 minutes to be used however necessary). In **Activity B, "Fishbowl" Discussion of Literary Work**, approximately 40 minutes are allotted to discuss the short story, and the remaining time is used by the teachers to write a brief summary of any revised insights into the story or observations about the group process that they may have after the discussion (these summaries are written in the assessment booklets). In **Activity C, Speaking of Language**, all of the teachers are given 10 minutes to prepare their oral presentation, approximately 5 minutes to give their presentation, and approximately 2 minutes to answer a follow-up question posed after their presentation. Much of the remaining time is taken by the teachers drawing their topic of presentation from a hat.

Both the teachers and the assessors were asked if they thought the time allotted for each activity was sufficient, too long, or not long enough. Approximately 58% (11 of 19) of the teachers thought the time allotted for each activity was sufficient, 32% (6 of 19) said it was not long enough, and 10% (2 of 19) said it varied according to the activity. None of the teachers said the time allotted was too long for any of the activities. The number of teachers specifying each activity as requiring more time is as follows:

	Requires More Time
Activity A, Responding to Student Writing	(4)
Activity B, "Fishbowl" Discussion of Literary Work	(1)
Activity C, Speaking of Language	(3)

In fact, five of the nineteen teachers did not finish **Activity A, Responding to Student Writing** in the time allotted. One of the four teachers who wanted more time for the activity explained:

> *I felt hurried to read, respond thoughtfully, and answer all questions on Activity A. Although I recognize the need for speed in responding to a class' worth of writing (30+ students), I felt the questions and the situation created a need for more time.*

One teacher who did not finish, however, considered the one- hour time limit to be sufficient because "teachers will realistically not be able to spend 15 minutes responding to each student's paper." Nevertheless, this teacher also admitted that "for teacher assessment purposes, I was unable to write all I wanted to show all that I was thinking."

Of the three teachers who wanted more time for **Activity C, Speaking of Language**, one called the activity "deep," and said it required "much more comprehension/analysis/synthesis" than the other activities. Another teacher commented,

> *In Activity C, preparing a speech in 10 minutes was quite difficult--and delivering it in 5 was nearly impossible with any references to the readings. I may have saved my nerves if I had prepared an outline for each topic at home.*

The third teacher who advocated more time for **Activity C** also suggested the idea of preparing an outline for each topic prior to the activity, especially as an alternative to allotting more time to prepare the speech at the assessment center.

Finally, the one teacher who wanted more time for **Activity B, "Fishbowl" Discussion of Literary Work** remarked,

> *I would have appreciated more time for Activity B because just about the time we were comfortable with each other our time was up.*

As for the assessors, because they did not administer two of the three activities, they were only asked to comment on the time allotted for the activity they did administer (i.e., **Activity B or C**). All four assessors thought the time for their activity was sufficient. One assessor explained why the time allotted was particularly suitable for **Activity B, "Fishbowl" Discussion of Literary Work:**

> *The discussions seemed to peak about the 30-minute mark. However, sometimes the "extra" 10 minutes allowed candidates who had started slow to recover. Groups in which one or two members are an impediment to the group process benefit from the full 40 minutes [of discussion]. Quieter, less assertive people seem to need some extra time to figure out how to cope with more vocal and assertive but less insightful candidates.*

Not all assessors perceived the discussions as peaking at 30 minutes, however. In fact, the assessor who was paired with the assessor quoted above remarked that "some groups were surprised that 40 minutes had passed so quickly when time was called". Nevertheless, as was discussed in the section, "Fairness Across Groups of Teachers," teachers differ in the ways they choose to articulate their skills and knowledge in a group discussion, and thus enough time needs to be allowed in the activity for these differences to be identified.

Two assessors also offered comments on the time allotted for **Activity A, Responding to Student Writing**. Both assessors seemed to be responding to the fact that some teachers did not finish the activity and/or gave brief answers to some of the questions. One assessor suggested that the teachers "be given a time warning 1/2 way so that they address both papers equally." The other assessor commented,

> *For some candidates with brief responses on Part II, I wasn't sure if they ran out of time or didn't have much to say....Maybe more time should be allowed for the activity.*

Based on our observations of the activities, and on the teachers' performance on each activity, FWL staff tend to agree with the teachers and assessors who consider the time allotted for **Activities B and C** to be sufficient. In response to those teachers who advocated directing the teachers to construct ahead of time an outline for each topic question in **Activity C, Speaking of Language** in order to be prepared at the assessment center, we agree that this is a possibility to be considered. However, because the activity is designed to measure a teacher's skill in delivering "impromptu" oral presentations--of the sort that might be given at teachers' or parents' meetings in response to audience questions--it seems that directing teachers to prepare ahead of time for the presentation (or giving them more time at the assessment center to prepare) would somewhat invalidate the "impromptu" nature of the activity. (In fact, for a truly impromptu presentation, consideration should be given to not providing the topic questions ahead of time with the set of research articles as this allows teachers to prepare an outline for each topic question if they choose.)

In response to the concerns raised by some of the teachers and assessors about the time allotted for **Activity A, Responding to Student Writing**, FWL staff notes that seven of the 19 teachers did not clearly pass the activity, and four teachers thought more time should be allotted for the activity. FWL staff recommends that consideration be given to extending the time for the activity by at least 10 to 15 minutes. Although we agree that teachers in practice do not have the luxury of unlimited time when responding to student writing, we think teachers should have adequate time when evaluating student writing for colleagues. Thus, the time allotted for the first part of the activity (i.e., responding to student writing) could remain the same, while more time could be added to the second part (i.e., evaluating the student text).

Clarity of the Rating Forms and Process

The rating process for the three assessment center activities was described briefly in the introduction and in the "Scoring" section. To recap, using a four-point scale, the assessors holistically rated the teachers' performances for each activity on three levels: (1) according to specific criteria listed under an evaluation category, (2) the evaluation category, and (3) the overall level. Each of the assessors was asked about their experience in rating the teachers for **Activity A, Responding to Student Writing** and for whichever activity they administered (i.e., **Activity B, "Fishbowl" Discussion of Literary Work** or **Activity C, Speaking of Language**). Their responses are discussed below.

Activity A, Responding to Student Writing. The assessors were first asked if they had any difficulties evaluating (a) Part I--the teacher's responses to the student writing, and (b) Part II--the teacher's analysis of the student text. The two new assessors said they had difficulties evaluating both parts. Of the two experienced assessors, one had difficulties evaluating Part I, and the other had difficulties evaluating Part II. Thus, three-fourths of the assessors had difficulties evaluating each part of the activity.

When asked to describe their difficulties and make suggestions for improvement, the three assessors who had trouble with Part I all referred to the very first part of the activity which requires the teacher to select, from eight possibilities, their purpose(s) in responding to the student writing samples. These possibilities are as follows:

a) Establish myself as a "friendly" audience
b) Inform the writer of problems for future work
c) Inform the writer of strengths s/he demonstrates
d) Demonstrate the primary criteria I'll be using to evaluate student writers, such as the importance of correctness, principles of organization, use of details, or other concerns
e) Guide revision of this particular piece of writing
f) Establish myself as an authority on good writing
g) Establish myself as an interested reader
h) Other (please explain)

The teachers are told in the directions for this part of the activity that their responses to the student writing samples will be evaluated in the context of their stated purposes. This, however, tended to present a problem for the assessors.

One assessor, for example, had difficulty rating teachers who marked purposes that she felt were inappropriate for the context of the assessment (i.e., the first writing assignment of the year):

> *The context obviously called for few to no corrections but candidates were allowed to choose that approach. Though it's poor pedagogy [on their part], I felt it unfair to penalize them.*

Another assessor said that, "based on candidate's choice of purpose," she was "sometimes not certain" how to rate the teacher's performance on two of the five criteria listed on the response form for Part I. A third assessor remarked,

> *Some of the purposes seem distinctly easier to carry out for new teachers (e.g., "friendly audience").*

Although only three of the assessors reported difficulty with Part I, all four assessors offered some suggestions for improving Part I. These suggestions were as follows:

- Include an item on the response form that allows the assessor to indicate whether or not the teacher's comments to the student are consistent with his/her purpose.

- To provide easy reference for the assessor, have a place on the response form that indicates what the teacher's purposes are.

- Improve the symmetry between the evaluation criteria listed on the response form and the list of purposes.

- Combine the purposes (a) and (g) into one statement.

In view of the problems experienced by the assessors, the suggestions made, and the earlier comments by Sharon Nelson-Barber which note the likelihood of different teachers choosing different purposes, it seems imperative that this part of the assessment (i.e., "Purposes in Responding") be revised. Although the "Purposes in Responding" seems to have been designed with the purpose of helping establish a "context" particular to each teacher against which evaluators might more accurately judge the teachers' comments, the assessment's design did not go far enough so that this purpose could actually be realized. FWL staff strongly recommend that the assessor's first three suggestions made above be followed, and that all of the purposes listed as well as the scored responses from this year's pilot test be reviewed and discussed to address the question of whether teachers who mark some purposes over others tend to get higher ratings.

For Part II, all three assessors who reported difficulties evaluating the teacher's analysis of the student writing samples described the same difficulty. The difficulty was evaluating the teacher's performance according to two of the five criteria listed on the response form. These two criteria--"B. Identifies effective features of the text," and "C. Identifies problematic features of the text."--actually require the assessor to rate five different dimensions each: content, structure, development, style/voice, and mechanics. The difficulties in rating came from these detailed criteria. Explained an assessor:

> *Part II, B and C were the most difficult to rate and come to an agreement on. Assigning a number value for each item was difficult since many of these items affect each other or overlap.*

In response to this problem, all three assessors suggested the same solution, described by one assessor as follows:

> *Instead of responding in such detail--respond for category, [and] leave listed features for assessors to comment on.*

In other words, for the two criteria, the ratings along each of the five dimensions would be eliminated; instead, the assessors would give one general rating for each criterion, and would use the five dimensions as guides to arrive at each rating.

Despite the difficulties mentioned above, none of the assessors had difficulty giving an overall rating to the teacher's performance on **Activity A, Responding to Student Writing.** One of the two new assessors who had difficulties evaluating both Part I and II, explained why she had no difficulty giving an overall rating:

> *If a student's paper was handled with intelligence and interest, taking into account the context, it was easy to weigh that heavier than any flaws in analysis for peers unless the errors or omissions were gross.*

Judging from her comment, it would appear that this assessor weighted Part I's rating(s) more heavily than Part II's. Although assessors were not instructed to do this, they were also not instructed not to do this. It would be interesting to know if the other assessors used the same or a similar process in arriving at their overall ratings. Interestingly, interrater reliability was the highest for the overall rating of this activity than for any of the other activities. But if this process was used, it raises some questions: Is this weighting desirable? If not, is it unavoidable? Since there are only two subsets, how does one arrive at a holistic rating if one subset is rated lower than the other? These questions can not be answered here, but in revising the rating form and process for **Activity A, Responding to Student Writing**, they should be explored.

Activity B, "Fishbowl" Discussion of Literary Work. Because only two of the four assessors scored this activity, only they were asked to answer questions about the scoring of this activity.

Of the two assessors, neither had difficulty evaluating the teacher's responses during the activity, and neither had difficulty evaluating the teacher's group process skills. One of the assessors, however, expressed difficulty with evaluating the teacher's interpretive skills. This assessor focused on two problems. The first was with the criterion, "Offers thoughtful and sound interpretive insights," one of five criteria listed under the evaluation category, Interpretive Process. The assessor described the problem as follows:

> *"Offers thoughtful and sound interpretive insights" -- this phrase still puzzles me. I can disagree with an interpretation that I find thoughtful. I don't see how I could disagree with an interpretation I considered sound. So what do I do with a thoughtful interpretation I disagree with?*

The second problem, as described below by the assessor, is less clearly defined, but raises an important question worth being considered:

> *I am also struck by the mullers and the formulators, those who put their finger on ambiguity and those who resolve it. Does the assessment reward glib formulators more than patient, tenacious mullers? I don't know.*

Indeed, perhaps the question is not whether or not the assessment rewards different styles of interpretation, but rather do the assessors? That is, when giving a rating, does the assessor consciously or unconsciously favor a particular style of interpretation or way of working in a group (i.e., group process skills)? Looking at the ratings again for **Activity B**, the answer would seem to be yes--at least for interpreting group process skills. Under the evaluation category, Group Process, six of the 19 teachers received two different ratings (e.g., a "3" and "4," or a "2" and "3") from the two assessors, indicating the likelihood that the two assessors were operating with different biases.

As with **Activity A**, the assessors for **Activity B** did not express any difficulties in giving an overall rating for the activity. One of the assessors did note, however, that the "overall rating of 3 covered a wide range."

Both assessors were also asked how frequently they used the teacher's log and the summary notes written at the end of the activity to aid in the evaluation of the teacher's interpretive skills. One assessor never used either item, but commented,

> *I think these could be very important for marginal candidates or for resolving discrepancies between assessors' scores.*

The other assessor reported using both items "for some teachers," and described the items as "very important" because,

> *When I was unsure if a candidate read and understood "The Lesson," I could check the log to verify both comprehension and interpretation.*

Two other comments were made related to the rating process for **Activity B**. One referred to the space on the response form reserved for comments. The assessors generally used this space to write down notes as the teachers were discussing the short story. These notes, however, were often not understandable to anyone who had not observed the activity. One assessor commented on the difficulty she had writing understandable notes, and offered a suggestion for improvement:

> *I had some difficulty in organizing my notes into comments that could be understood by a reader who had not observed the activity.*
>
> *Perhaps notes and comments should be written on a separate piece of paper. Divide the paper into two columns--one for notes taken during the activity, the second for commentary that would explain the relationship of the notes to the ratings.*

Assuming that it is important for the ratings to be supported by some sort of legible and understandable evidence, especially when the rating is a negative one, FWL staff believes this assessor's suggestion should be strongly considered. At the very least, any future training for assessors should include instruction on how to write their comments in an appropriate manner.

Finally, an important recommendation was made by one of the assessors, possibly as a result of her experience on the last day of the pilot test. On that day, because there was only one group of four teachers being assessed, the two assessors of **Activity B, "Fishbowl" Discussion of Literary Work** were able to view the administration of **Activity C, Speaking of Language** after they had finished and scored their activity. After both activities were completed, FWL staff heard the assessors commenting on the differences in the teachers' performances in the two activities (i.e., **Activity B and C**). Thus, FWL staff agrees with the following assessor's recommendation and contention:

> *Continue to have separate assessors for Activities B and C. If I had observed candidates' participation in C, it could have influenced my ratings for B.*

Activity C, Speaking of Language. Of the two assessors who administered and scored this activity, the experienced assessor reported no difficulties with evaluating the teacher's performance for this activity, while the new assessor experienced two difficulties.

First, the new assessor experienced difficulty evaluating the teacher's responses during the activity (i.e., while the teacher was giving his/her oral presentation). As she explained,

> *It was difficult to attend to content and take notes (as in college lectures).*

The assessor also added that the difficulty was worse for some presentations than others because some of the presentations did not correspond as well to the scoring criteria listed on the response form for the activity.

This assessor also experienced difficulty evaluating the teacher's skills in planning the presentation. In particular, the assessor had difficulty rating the teacher on two of the five criteria listed under the evaluation category, Plan of Presentation. These two criteria, "Communicates clear central idea or question" and "Clarifies issues with analysis or reasoning," were especially difficulty to rate, said the assessor, under the following circumstances:

> *If a teacher's ideas--however clear or well analyzed--did not fit the research or take into account the totality of the issue (e.g., "canon vs. multi-cultural"), I didn't know how to assess them.*

The assessor seems to be saying that if the teacher's presentation communicated a clear central idea or question and clarified issues with analysis or reasoning, but did not reflect the set of research articles to be read in preparation for the activity or perhaps only addressed a small part of a very large issue, then the assessor was not sure how to rate the teacher on the criteria named above. Or, in other words, how should an assessor rate a teacher who gives a good presentation, but does not really address the issue? This question is not addressed in the present training design, but should be in the future.

Although both assessors said they had no difficulty giving an overall rating for the teacher, it should be noted that six of the 19 teachers received two different overall ratings (e.g., a "3" and "4" or a "3" and "2") from the two assessors for this activity. In the comment below, the experienced assessor makes an important observation about the scoring differences between him and his fellow assessor:

> *We consistently agreed on our evaluations of the candidates' performances; the only time we had a major disagreement on score, we discovered, upon discussion, that we had both seen the same shortcomings in the candidates's presentation, but disagreed on how much these shortcomings should lower his score.*

Thus, because the scoring is a holistic process versus being anchored to samples of performances, different reactions to the same data can result in different scores. Perhaps if more examples were provided in the training to illustrate ratings, scoring differences could be reduced.

When asked how frequently they used the teacher's reading log to aid in their evaluations of the teacher's content and organization skills, both assessors said they used it "for some teachers." Commented one assessor,

> *For those whose presentations were less than very strong, the log served as an additional source of information that might help the candidate's score.*

The other assessor also valued the reading log and suggested that it be rated under the subset, Plan of Presentation.

One other comment was made by an assessor about the format of **Activity C**. The assessor suggested adding one or two articles about adolescent literature to the set of articles in the handbook for the activity. The addition of the articles would serve two purposes: (1) new questions could be created to serve as topics of presentations, and (2) the addition of more questions would make it possible for each teacher to draw two questions, selecting one and discarding the other. As the original assessment activity included questions about adolescent literature--but no corresponding articles--FWL staff think the addition of such articles and questions would enhance the activity, as would the revised format of allowing each teacher to select one question after drawing two.

In conclusion, revisions need to be made to the rating process and forms for all of the three assessment center activities, but particularly to the process and form for **Activity**

A, Responding to Student Writing. In addition, based on the assessors' comments and the scoring results, consideration needs to be given to greatly reducing the probability of assessors interpreting the same data differently. That is, there needs to be more consensus as to the kind of performance that is represented by each point on the rating scale. Providing the assessors with a well-written scoring manual with numerous examples during training could help address this problem. Training and a scoring manual could also address the way in which assessors should write their comments to support the ratings they make for each activity.

Cost Analysis

Administration and Scoring Cost Estimates

The Secondary English Assessment: Assessment Center Activities is administered in an assessment center format. The current structure of the activities are such that four candidates can be administered the three activities in a half-day session using four assessors. Thus, it requires approximately one half-day assessor of time per half-day assessment for each candidate. An additional hour for preparation and finalizing as assessment is needed for each assessor. Using a rate of $20/hour yields an estimate of $100/assessment for administration and scoring of the three activities.

Training for this assessment was two days. Future assessments would require at least this amount of training and the training could be extended to three days. If we assume that each assessor would conduct 30 assessments each year for five years, we could distribute the costs for training an assessor over 150 assessments. Reimbursing assessors for three days training at $160/day or $20/hour would cost $480. Distributing the $480 over the 150 assessments adds approximately $3/assessment for training.

Other costs include those associated with telephone, duplication, postage, and travel where needed. Travel could be expensive in California unless regional assessments were used. A regional assessment would minimize travel costs. Estimating costs for these activities or ingredients would depend in large part on the manner in which the system was ultimately designed and how costs were apportioned. Using a figure of $30 per assessment for these activities would assume only minimal travel costs, based on our experience from the pilot testing. This is the same estimate that was used in the First Year Report on Pilot Testing.

These result in the following cost estimates for administering and scoring the Secondary English Assessment: Assessment Center Activities in a half-day assessment format:

Assessor Costs:	$100/assessment
Training Costs:	$3/assessment
Other Costs:	$30/assessment
Total Admin/Scoring	$133/assessment

Development and Pilot Testing Costs

The costs for developing all four activities of the Secondary English Assessment (i.e., the three assessment center activities and the portfolio activity) were $84,415 and are broken out by Cost Category in Table 7.5 which also includes costs for pilot testing. These development costs are the expenses for the assessment developer to deliver drafts for these activities to the CTC and SDE. The developer was building on prior work with these assessment activities and approaches; thus, future development costs would be more similar to these than if a new development effort was initiated. Additionally, approximately $45,429 were incurred for the pilot testing of these assessments with 19 teachers.

These provide samples of developmental costs that should be considered if a similar assessment were to be adapted for implementation.

Technical Quality

This section discusses the technical issues related to the three assessment center activities of the Secondary English Assessment--development, reliability, and validity.

Development

Although this assessment was developed during the period of May 17 to December 29, 1989, two important sources of information contributed to the preliminary design stages: (1) an August 1987 California State University workgroup, and (2) the piloting of the English 677 course at San Francisco State University during the spring semester of 1988. The workgroup created a comprehensive list of desired competencies for prospective English

TABLE 7.5

DEVELOPMENTAL AND PILOT TEST COSTS FOR THE SECONDARY ENGLISH ASSESSMENT

Cost Categories	Development	Pilot Testing
Staff-Salaries & Benefits	$21,219	$ 14,292
Consultants (Teachers, assessors, and other consultants)	37,210	15,491*
Travel (Consultants and staff)	5,103	4,044
Other Direct Costs (Site rental, phone, duplication)	4,000	2,085
Total Direct Costs	$67,532	$35,912
Indirect Costs	16,883	9,517
Total Costs	$84,415	$45,429

*These costs are those for developing the three activities pilot tested and reported here and a portfolio to be pilot tested this fall-winter. Pilot test costs are those for pilot testing these three activities.

teachers and proposed some alternative plans for assessing these competencies. The experimental English 677 course, titled "Performance in English," assessed selected performance abilities of credential candidates in English. Of the approximately 15 assessment activities piloted in the course, three were selected for inclusion in the Secondary English Assessment developed for the California CDE/CTC.

After the identification of the assessment activities, three groups of participants were identified and recruited to review the activities and/or to develop assessment materials and procedures for the activities. These three groups were: (1) five expert English educators involved in teacher training institutions, (2) eight veteran English teachers, and (3) 16 new teacher candidates who acted as subjects of the assessment activities and assisted in revising and refining the assessment package.

Upon completion of draft materials and procedures for each activity, a pilot test was conducted in an August 1989 workshop. The participants in the pilot test were the eight veteran English teachers and the 16 new teacher candidates. Based on the results of this workshop, revisions were made to the assessment materials and procedures. Further revisions were also made throughout the fall by the assessment developer and two of the veteran English teachers, and final revisions were made after a December 1989 meeting of the eight veteran English teachers.

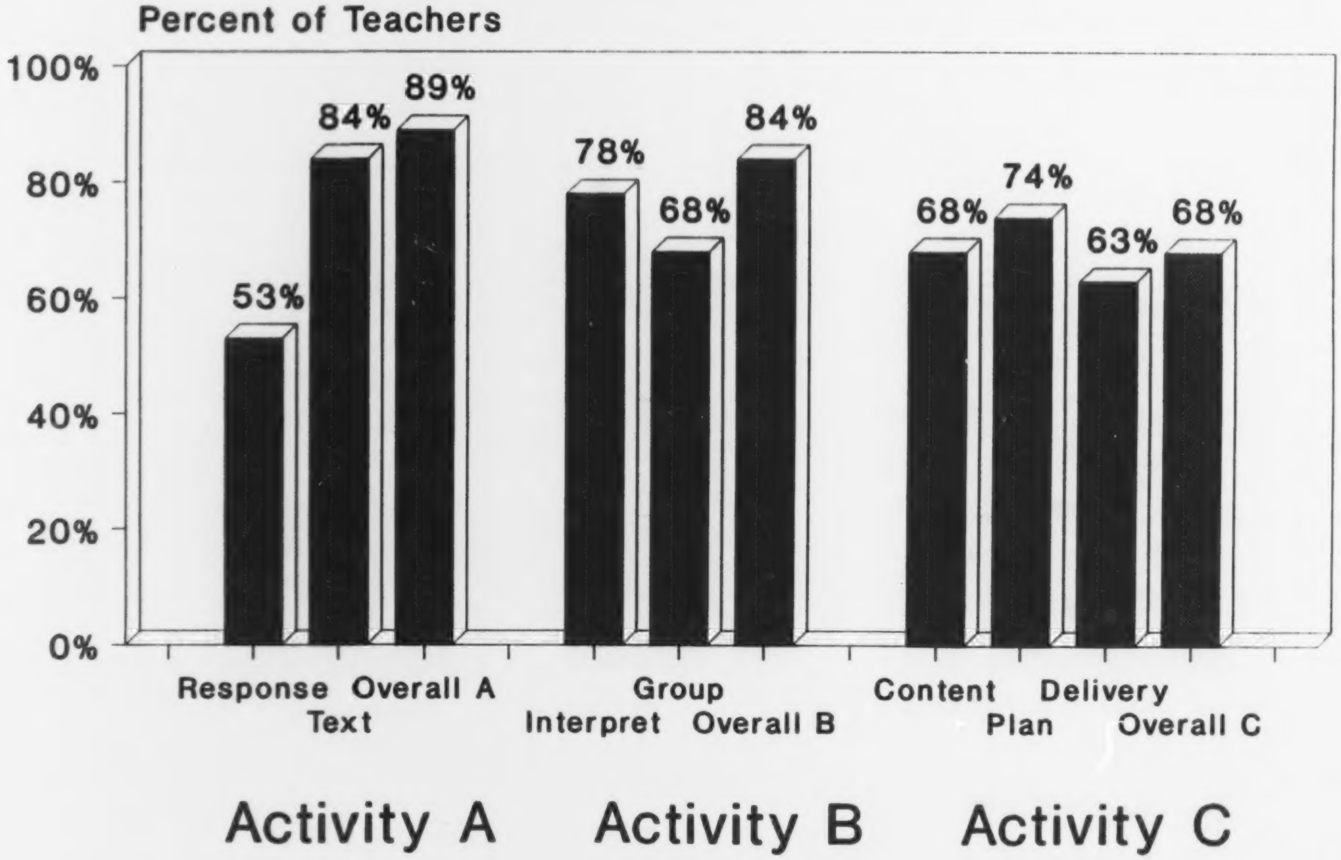
Reliability

The following analyses were performed on the pilot test data of 19 teachers. Interrater agreements were examined to assess the degree to which assessors were able to consistently judge candidates using the English-Language Arts Assessment scoring protocols. Internal consistency estimates were generated to assess the degree to which the variables or factors within each of the activities would form a measure and the degree to which the different activities related to each other and might form an overall assessment of a candidate.

Interrater agreements. The first measure of agreements among judges was obtained by comparing the number and percent of ratings in which assessors gave identical or different ratings. Figure 7.2 presents the percent of exact agreements for **Activities A, B and C**. They range from a low of 53 percent for **Activity A's** Overall Response Strategies to a high of 89 percent for **Activity A's** Overall Rating. The only variable on which raters differed by more than one point was on **Activity B's** Interpretive Process where two of 19

FIGURE 7.2

Percent Agreement Between Raters for the Secondary English Assessment Activities



teachers received ratings two points apart. This level of agreement on the three activities suggests that a fairly high degree of agreement has been achieved by the training and scoring associated with the pilot test.

Interrater correlations. Correlations between raters also serve as an estimate of interrater agreement. The correlations among rater pairs are displayed in Table 7.6. Correlations were calculated for each variable on each of the activities and for the Overall Rating on each activity. For each variable there are two measures. The first which is labeled **Rating (RT)** is the holistic rating given for the variable. The second which is labeled **Summed (S)** was obtained by summing the individual items for that variable to form a score. Two rater pairs rated **Activity A**: Responding to Student Writing; three rater pairs rated **Activity C**: Speaking of Language. The average correlations across rater pairs were calculated and are presented for these activities. For **Activity B**: Fishbowl Discussion of Literary Work, a single rater pair rated all 19 teacher candidates and this correlation is reported.

Using the average across rater pairs, the interrater correlations for holistic ratings range from **Activity A**: Responding to Student Writing's Overall Response Strategy where the correlation is zero (0) to **Activity C**: Speaking of Language's Presentation Plan where the average correlation was .96 for the holistic rating. There was no particular pattern for the holistic ratings and Summed Ratings to have higher or lower agreements among the raters. Thus, it appears that using only the holistic rating for each variable or omitting the holistic ratings for the subparts of each activity and summing the individual ratings are both viable approaches in terms of the degree of agreement that will be observed between raters. Holistic ratings could reduce the rating time.

The variability of these correlations reflect both random fluctuations due to the small numbers of teachers rated and a need to further refine and develop the rating system. Given the draft status of this assessment, these results suggest that the assessment and scoring systems could be developed to yield reasonable agreements among raters on these types of tasks or assessment.

The rating system allows for the raters to rate NA on those items for which they judge insufficient information was available to make a rating. Examining the degree to which rater pairs observing the same candidates rated the same items as NA also provides a measure of rater agreement. Across all items for the three activities, 32.3 percent of the ratings had one rater but not the other rate an item as NA. Thus, for about one third of the ratings one assessor but not the other judged there was not sufficient evidence on which

TABLE 7.6

CORRELATIONS BETWEEN RATERS FOR THE SECONDARY ENGLISH ASSESSMENT ACTIVITIES FOR HOLISTIC RATING (RT) AND SUMMED RATINGS (S)

Activity/Part	Types of Rating	Rater Pair			Averaged Pair Ratings
		1	2	3	
Activity A, Responding to Student Writing					
Response Strategies	RT	0	0		0
	S	.07	.26		.16
Analysis of Writer to Text	RT	NA	.65		NA
	S	.22	.49		.36
Overall Activity A	RT	.25	1.00		.90
	S	.25	.75		.47
N's		11	8		
Activity B, "Fishbowl" Discussion of Literary Work					
Interpretative Process	RT	.48			
	S	.55			
Group Process	RT	.53			
	S	.17			
Overall Activity B	RT	.74			
	S	.39			
N		19			
Activity C, Speaking of Language					
Content	RT	.65	.85	.91	.83
	S	.64	.50	.87	.71
Plan	RT	.56	1.00	1.00	.96
	S	.58	.71	1.00	.88
Delivery	RT	.51	.33	.82	.59
	S	.51	.33	.82	.59
Overall Activity C	RT	.69	.85	.59	.73
	S	.50	.85	.52	.74
N's		11	4	4	

to make a rating. An implication is that the training and ratings would be strengthened by using additional tapes of teacher candidates to provide examples on which raters can rate and discuss ratings including what constitutes sufficient evidence for making a rating.

Internal consistency of the tasks and assessment. Coefficient Alpha reliability estimates were calculated for the three different activities and their subparts by using the individual ratings on items within each subpart. The reliabilities for the activities and subparts are listed below:

Activity/Subpart	Reliability	
	Subpart	Total
A: Responding to Student Writing		*
Response Strategy	.77	
Analy. Writer & Text	*	
B: Fishbowl Discussion		.90
Interpretive Process	.86	
Group Process	.85	
C: Speaking of Language		.91
Content	.74	
Presentation Plan	.83	
Delivery	.78	

* Indicates that insufficient ratings were made due to the number of NA ratings.

These estimates indicate a relatively high degree of internal consistency within the subparts and total activity evaluations/ratings. A review of the data suggest that raters tend to assign the same rating to items within an activity and to the overall rating for the activity. An implication is that it might be possible to have raters simply assign ratings to the different subparts and not take the time to rate individual items within subparts. This would allow for providing some feedback on candidates' strengths and weaknesses within each of the activities and lessen the time for ratings from what is required if individual ratings are made for each item. The reliabilities for each activity suggest that the activities do form a measure in which an overall judgment or evaluation can meaningfully be made. If there was low or no internal consistency within the activities or their subparts, then it would call into question what is being measured and would undermine the interpretability of any composite or overall evaluation for the activities.

Intercorrelations among activities. Correlations among the three activities were calculated for the 19 teacher candidates and are reported below.

	A	B	C
Activity A: Responding to Student Writing	-		
Activity B: Fishbowl Discussion	.26	-	
Activity C: Speaking of Language	.58	.36	-

Given the relatively small N of 19, only the .58 correlation between **Activity A and C** is statistically significant. However, these correlations provide some support that although the activities are related there will be not be as strong a relationship across activities as there is consistency within each activity. Further support for this is provided when coefficient alpha is calculated as a measure of reliability across all activities. For the 19 candidates in the pilot test the internal consistency across activities was .67. This indicates some internal consistency across all activities but also provides tentative evidence that the activities measure somewhat different attributes of the teacher candidates' performance.

Validity of Agreement Through Group Comparisons

Differences in performances were examined for minority-nonminority, women-men, high school-middle school, urban-inner city-suburban, teachers and for the number of courses teachers had completed in the subject area. It was felt that this could provide at least preliminary glimpses of the assessment's difficulty for different groups. Some of these analyses that compare different groups have been discussed in earlier sections. The pilot test sample size and design were not constructed to provide information sufficient to provide stable estimates comparing differences among these groups. For example, some subgroups have as few as three teachers in them. Nevertheless, an examination of differences among groups provides some initial insights into the validity of this assessment. Table 7.7 contains a summary of the trends for the pilot sample of 19 teacher candidates. Appendix E provides the means, standard deviations and numbers of candidates from which these summaries were constructed. A plus (+) simply indicates that the mean or average for the first group

TABLE 7.7

TRENDS OF MEAN DIFFERENCES BETWEEN CANDIDATES WITH DIFFERENT CHARACTERISTICS FOR ACTIVITES AND EVALUATION CATEGORIES*

Activity	Gender Female/ Male	Preparation More/ Less Courses	Level of Teaching HS–Middle/ Jr.High	Teaching Location Suburban (Urban Inner City)	Ethnicity Non-Minority-Minority
Activity A, Responding to Student Writing					
Response Strategies	+	+	+	+	+
Analysis of Writer to Text	+	+	+	+	+
Overall Activity A	+	+	+	+	+
Activity B, "Fishbowl" Discussion of Literary Work					
Interpretative Process	+	+	+	+	–
Group Process	+	–	+	+	–
Overall Activity B	+	+	+	+	–
Activity C, Speaking of Language					
Content	+	+	+	+	–
Plan	+	+	+	+	+
Delivery	+	+	+	+	+
Overall Activity C	–	+	+	+	+
SUMMARY	10/10	8/9	9/10	10/10	6/10

*Entries reflect the direction of the mean differences for the different candidates. For example, in the activity and the evaluation category Responding to Student Writing, Response Strategies, the average or mean of female teachers in the pilot test was greater than the males. These do not generally represent statistically significant differences and due to small N's no tests of significance were calculated.

was greater than that for the second group. For example, the pluses under the Female-Male column indicate that for 9 of 10 variables the female's average was greater than the males.

Trends in the table provide some preliminary evidence that on this assessment and pilot test sample:

- females' averages were greater than males on 9/10 evaluation categories;
- teachers with more preparation/courses had averages that were greater than those with less (i.e., 0 or 1 course) on 9/10 scores;
- high school teachers' averages were greater than junior/middle school teachers' on 10/10 evaluation categories;
- teachers in suburban settings had averages that were greater than those in urban/inner city settings on 10/10 evaluation categories; and
- nonminority teachers' averages were greater than minority teachers' on 6/10 evaluation categories.

If these trends were to hold for larger, more representative samples, some of these trends would be encouraging evidence for the "validity" of the assessment, whereas others would provide less encouraging findings. For example, if teachers with more preparation courses perform superior to those with fewer, this would provide some positive evidence that the assessment does differentiate among those with greater and less knowledge/preparation. Similarly, it would be desirable for the assessment to minimize any adverse impact on minority teacher candidates. Thus, although the difference between minority and nonminority teachers was less than others (e.g., preparation or teaching location), a sample of three prohibits drawing any conclusions about how minority teachers would perform on this assessment.

The above comments address ways in which group performances can contribute to evaluating the appropriateness and difficulty of the assessment. However, the results also can provide information that will be useful for teacher preparation, training and recruitment. For example, it is not encouraging that urban and inner city teachers score lower than others. This finding provides further support that the urban and inner city schools may very well not be attracting or securing as strong a new teacher force as suburban schools. The trend for middle and junior high school teachers to perform less well than high school teachers could imply a need for strengthening the content-specific preparation of these teachers.

Content validity. The content validity of this assessment rests largely in the role that teachers and English educators had in the development, and the analyses of the match of the assessment to the model curriculum and teaching standards which supplies evidence that the assessment's contents have validity with respect to current and emerging content. These have been described earlier and implications for further development are described in the following section.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the three assessment center activities of the Secondary English Assessment. The section presents information in the areas of administration, content, and format, and concludes with a brief summary.

Administration of Assessment

Each of the three assessment center activities of the Secondary English Assessment is approximately 50 minutes long. As administered in this pilot test, the first activity was scored in the afternoon, after the teachers had completed all of the assessment activities. The second and third activities were scored during and immediately after they were administered. Thus, for this pilot test, the administration and scoring of the three assessment center activities of the Secondary English Assessment required approximately five hours a day per assessment.

Based on our experience, the following factors seem to be key to smooth implementation of the Secondary English Assessment: Center Activities (or any assessment that includes similar assessment center activities):

- recruitment of assessors who are experienced English teachers, who have had experience in formal writing assessment programs involving holistic scoring of writing samples, and who are knowledgeable about different teaching styles, interactive styles, and patterns of communication;
- availability of appropriate assessment center facilities (e.g., two rooms for assessment);

- development of clear orientation materials for teachers which include descriptions of each assessment activity, the criteria by which the teachers will be evaluated, and all preparation materials needed to be read/completed before assessment day; and

- development of procedures to collect and store assessment materials for each activity such as assessment booklets, evaluation response forms, and completed preparation materials (e.g., reading logs).

Another key factor to the smooth implementation of assessment center activities such as those of the Secondary English Assessment is assessor training. Good assessor training serves to familiarize the assessor candidates with the content of the activities, as well as how to administer and score the activities. Although all of the assessors described the training for this pilot test as "very good," they also made some suggestions for improvement. Based on the assessors' comments, our observation of the training, and on performance data from the assessment activities, we believe the training could be improved by following these recommendations:

- Development of an assessor handbook to serve as a guide for assessors when administering and scoring the activities. In particular, the handbook should include a complete description of the scoring process and specifically provide concrete examples whenever possible of (1) the distinctions between rating points, and (2) the way in which comments should be written on the scoring response sheets.

- Extension of the training by one half to one full day in order to address more thoroughly the material covered in the assessor handbook (e.g., the details of the scoring system).

Following the above suggestions should greatly facilitate the administration of the assessment center activities.

Assessment Content

Based on the observations of FWL staff, as well as information from assessors, teachers, our consultant on cultural diversity, and the assessment documentation (i.e., the scoring response forms for each activity), the following conclusions are offered about the content of the Secondary English Assessment's three assessment center activities.

- Congruence of the three assessment center activities with the English-Language Arts Framework is weak. **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language** would especially have to be revised in order to achieve strong congruence. In particular, the two activities would need to be changed so that there is a greater focus on a teacher's skill in responding to or developing students' abilities versus a demonstration of skill in activities that have only an indirect relationship to teaching students.

- Coverage by the Secondary English Assessment: Center Activities of the California Standards for Beginning Teachers is also weak. Not one of the standards is fully covered by any of the activities, and not one standard is directly addressed by **Activity B, "Fishbowl" Discussion of Literary Work** or **Activity C, Speaking of Language.** Although most of the standards are addressed by **Activity A, Responding to Student Writing,** they usually are done so in an indirect or limited way.

- Based on the teachers' and assessors' comments, the three assessment center activities seem to be job-related, although **Activity C, Speaking of Language** seems less so, and **Activity A, Responding to Student Writing,** more so. In addition, **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C** are probably best described as indirectly job-related because they do not directly assess in any way a teacher's teaching skills.

- Of the three activities, the teachers had the least difficulty with **Activity B, "Fishbowl" Discussion of Literary Work,** with almost 90% of the teachers passing. **Activity A, Responding to Student Writing** and **Activity C, Speaking of Language** were passed by 63% and 68% of the teachers respectively, suggesting that beginning teachers may have had less opportunity to acquire the skills and knowledge measured by those activities than **Activity B.**

- Teachers and assessors thought the assessment is fair to teachers of different grade levels. However, analysis of the performance data seems to suggest that junior high/middle school teachers may be less well prepared than senior high teachers for **Activity A, Responding to Student Writing** and **Activity C, Speaking of Language.**

- When asked to comment on how well the assessment activities address a beginning teacher's ability to work with diverse students, the assessors commented favorably about **Activity A, Responding to Student Writing** but found **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language** to be less suitable. Our consultant on cultural diversity pointed out, however, that assessors need to be are familiar with the current research on students of different racial/ethnic groups as it pertains to the topic of the assessment activity (e.g., student writing) for **Activity A** to be a fair assessment.

- With regard to the question of the assessment's fairness across groups of teachers (e.g., different ethnic groups, different language groups), the majority of teachers and assessors responded positively. The issue of fairness is also largely dependent upon the assessors' knowledge of the possible teaching practices and styles of different groups of teachers.

- The performance data indicates that, with regard to different groups of teachers, females tended to receive higher overall ratings than males for **Activity A, Responding to Student Writing** and **Activity B, "Fishbowl" Discussion of Literary Work**, and lower for **Activity C, Speaking of Language.** Also, those teachers who described themselves as teaching in suburban locations tended to receive higher ratings in all of the evaluation categories for all three activities. The performance ratings of our small sample of minority teachers (3) were mixed (i.e., some higher and some lower) in comparison to those of non-minority teachers.

- While a majority of teachers and all of the assessors think the three assessment center activities are an appropriate way of assessing skills in teaching English classes, many of the teachers and some of the assessors reject the notion that these activities are an appropriate way of assessing general teaching skills. In particular, the teachers faulted the assessment center activities for not requiring any teaching.

Assessment Format

Although the format of each of the three assessment center activities is distinctly different, the format of each has these two features in common: (1) preparation work to be completed before the assessment day, and (2) a performance-based activity administered at an assessment center. Based on comments by teachers, assessors, and FWL staff, the following conclusions and recommendations are offered regarding the format of the preparation work and of the performance-based activities:

- Although the majority of teachers read the Orientation Handbook carefully and were satisfied with the materials presented, almost one third of the teachers said they had difficulty with the preparation work required for **Activity C, Speaking of Language**. The problem cited most often was that all or some of the research articles for the activity were difficult to understand or not interesting--a difficulty which may indicate a lack of experience in reading research articles.

- The time allotted for each activity (approximately 50 minutes) was deemed sufficient by the majority of teachers. Consideration could be given, however, to extending the time allotted for part two of **Activity A, Responding to Student Writing**, as this activity was not finished by approximately 25% of the teachers.

- Based on assessor comments, the teachers' performance data, and an examination of the completed rating response forms, revisions need to made to the rating process and forms for all three of the assessment center activities. For each of the following activities, we recommend the following:

  Activity A, Responding to Student Writing

  -- Revise the scoring process and response form so that the teachers' answers to the "Purpose in Responding" part of the activity are taken into account on the scoring response form.

-- Revise the scoring response form so that there is greater symmetry between the list of purposes and the evaluation criteria on the response form.

-- Review the scored response forms from this year's pilot test to address the question of whether the teachers who mark some purposes over others tend to get higher ratings.

Activity B, "Fishbowl" Discussion of Literary Work

-- Address the question of whether the scoring process, as it is presently constructed, allows the assessors to consciously or unconsciously favor a particular style of interpretation or way of working in a group.

-- Consider revising the rating process for the activity so that an additional form is used which allows assessors to write their notes and comments on one side of the form, and on the other they can give commentary that explains the relationship of their notes to the ratings.

-- Continue to use separate assessors for this activity and **Activity C, Speaking of Language** so that a teacher's performance in one activity does not influence his/her ratings in the other activity.

Activity C, Speaking of Language

-- In any future training, address the question of how to score a teacher who gives a good presentation but does not really address the issue presented in the reading material.

-- Consider adding additional articles and corresponding questions so as to expand the set of questions used for the oral presentations, making it possible for each teacher to draw two questions (instead of one), and then to select one and discard the other.

- For all activities, consider having assessors only give the holistic rating for the evaluation categories and the overall activity, using the evaluation criteria listed under each category as prompts and guides for summarizing and making comments to support the ratings and highlight candidates strengths and weaknesses.

Summary

While all three assessment center activities of the Secondary English Assessment are innovative and strongly performance-based, only **Activity A, Responding to Student Writing** assesses skill and knowledge that is directly related to the teaching of students in an English class. As they are now constructed, the content and format of the other two activities, **Activity B, "Fishbowl" Discussion of Literary Work** and **Activity C, Speaking of Language**, seem better suited to staff development purposes, although revisions could possibly be made to the activities so that they more directly relate to the teaching of students.

CHAPTER 8:

SECONDARY ENGLISH ASSESSMENT: PORTFOLIO ACTIVITY

Developed at San Francisco State University, the portfolio activity is one of four activities constituting the Secondary English Assessment (for a discussion of the other three activities, see Chapter 7). While the first three activities were designed to be administered at an assessment center during a half-day period, the portfolio activity is completed over a three-month period by the teacher in his/her classroom.

The portfolio assessment evaluates a teacher's skills in three areas: planning and implementing a unit, responding to student work, and reflecting upon his/her experience in teaching the unit to gain insight for further teaching. For the assessment, the teacher plans and conducts a three- to six-week teaching unit in which the classroom activities are unified by a single focus (e.g., a novel, a particular genre, a set of skills). To document the teaching activities, the teacher compiles a classroom portfolio which consists of the following tasks, also referred to as components:

(1) **Outline of Unit Plan**--The teacher outlines the unit plan, including written descriptions of the following: (a) context, (b) unit focus, goals, and rationale, (c) the sequence of activities, and (d) multicultural perspectives.

(2) **Weekly Log**--The teacher keeps a weekly log as a record of significant events, anecdotes, insights, and questions that convey the flavor of activities over the duration of the unit.

(3) **Materials and Assignments Given to Students**--All materials and assignments that are part of the teaching unit should be included in the portfolio. Assignments given orally are to be paraphrased in writing.

(4) **Samples of Student Work, with Teacher Responses**--The teacher provides samples of students' work, with the teacher's written responses, including (a) the work of one student for the entire unit, and (b) a mix of student responses for one activity.

(5) **Student Evaluations**--The teacher collects written evaluations from the students, either of the entire teaching unit or of one major activity.

(6) **Reflective Essay**--The teacher writes an essay that is a reflection upon the unit taught. The teacher is expected to address such questions as, What did you learn about this unit? Your students? Yourself as a teacher? The teacher is also expected to reflect upon his/her experience as it relates to the teaching and learning of English in a multicultural society.

The scoring system for the portfolio assessment is complex and varied. The portfolio is evaluated using a scoring response form which is divided into six parts. Each part represents a different skill or teaching competency, and the teacher's performance related to each skill is evaluated at two levels: (1) according to specific criteria listed for the skill/competency, and (2) an overall rating at the skill/competency level. For example, Part I of the scoring response form (see Chart 8.1) evaluates the teacher's planning abilities. At the first level of evaluation, the scorer assigns a rating for each of five criteria using a three-point scale: adequate (+), marginal (-), and missing or inadequate (M). To give an overall evaluation of the teacher's planning abilities, the scorer uses a four-point scale, with a rating of "4" indicating a very strong performance and a rating of "1" being a very weak performance. Some parts of the scoring response form use both the three-point and four-point scales, and other parts use only the four-point scale. All of the ratings are made in a holistic manner and are not interdependent.

The administration, the content, and the format of the portfolio activity of the Secondary English Assessment are discussed below. The content and format sections of the report contain information from the teacher and assessor evaluation forms, as well as information and analysis of scoring results. Following these three sections are sections on cost analysis and technical quality. The chapter concludes with an overall summary together with recommendations for further steps in exploring the feasibility and utility of assessment activities such as this in California teacher assessment.

Administration of Portfolio Activity

Beginning with an overview of the administration of the portfolio activity, this section provides information on the following: logistics (e.g., scheduling the activity, recruiting and training scorers), security, assessors and their training, teacher and FWL staff perceptions of the portfolio administration, and scoring (including characteristics of scorers and their training).

Response Form D Chart 8.1

Candidate ______________________ Evaluator: __________ Date____/____/____

For sections IV, V, and VI, circle appropriate score:

4 = Available evidence suggests definite strength in this area;
3 = Evidence suggests some strength in this area;
2 = Evidence suggests minor weaknesses in this area;
1 = Evidence suggests serious weakness in this area;
N.E. = No evidence on which to judge this area.

IV. Portfolio record of the **UNIT AS TAUGHT** reveals these general pedagogical abilities:

Strong.........Weak		
4 3 2 1 N.E.	1.	Shows flexibility and adaptability when needed.
4 3 2 1 N.E.	2.	Shows understanding of student attitudes and feelings.
4 3 2 1 N.E.	3.	Shows ability to reconcile conflicts between demands of context and goals of instructions.
4 3 2 1 N.E.	4.	Appears to use greater balance of class time for high-interest, student-centered activities, with a minimum of inappropriate, irrelevant or unchallenging "busy work".
4 3 2 1 N.E.	5.	Shows ability to **sequence activities** in a way that enhances learning.
4 3 2 1 N.E.	6.	Evaluation methods (measures/activities) are appropriate for determining student outcomes, class progress toward goals of unit.
4 3 2 1 N.E.	7.	Evaluation criteria are clear to student.
4 3 2 1 N.E.	8.	Evaluation methods(measures/activities) are themselves vehicles for promoting thoughtfulness and further learning.

Comments:

Overall Evaluation of General Pedagogical Abilities (circle one):

Strong..........Weak
4 3 2 1 N.E.

Overview

Nineteen secondary English teachers participated in the August, 1990 assessment center activities of the Secondary English Assessment and agreed to complete the portfolio activity between September and December 7, 1990. By the December 7 deadline, however, only 12 portfolios were submitted, and thus the deadline was extended. By January 31, 1991, the final deadline, 16 portfolios had been received. One teacher never attempted to participate in the portfolio activity, another teacher was unable to complete the portfolio due to health problems, and another teacher dropped out without explanation.

As shown in Table 8.1, of the 16 English teachers who completed portfolios, the majority were Caucasian (non-Hispanic) females teaching at the high school level. Almost an equal number of teachers represented schools in northern and southern California; three teachers represented schools in the central valley. A little over one third of the teachers were teaching in inner city schools. Seventy-five percent of the teachers were participating (or had participated) in CNTP-sponsored teacher support projects.

Logistics

The administration of this pilot test entailed numerous logistical activities, including contacting the teachers identified to participate in the activity, making follow-up phone calls, arranging for the mailing of the portfolios, recruiting and training scorers, and acquiring evaluation feedback from the teachers and the assessors. In addition to these activities, there were two other important logistical activities relevant to this pilot test: recruiting trainers and developing the training for scorers, and some revising of the developer's original scoring response forms.

Contacting the identified teachers. The teacher sample for this pilot test was the same group of teachers who had participated in the pilot test of the assessment center activities of the Secondary English Assessment. The sample included Project and non-Project teachers, all of whom were offered $300 to participate in the assessment center activities and to complete a portfolio. Each of these teachers was contacted in September by a FWL staff person designated as the Portfolio Contact Person (PCP). The PCP asked the teacher to name the focus and length of the teaching unit to be taught and the dates s/he planned to teach it. The PCP then sent the teacher a letter confirming the information regarding the unit's focus, length, and dates.

TABLE 8.1

PILOT TEST PARTICIPANTS
PORTFOLIO ACTIVITY
SECONDARY ENGLISH ASSESSMENT
(Number of Teachers = 16)

Location	No. of Teachers		Teacher Characteristics
	Project	Non-Project	
Northern California	4	2	14 Caucasian, non-Hispanic; 1 Hispanic; 1 Asian or Pacific Islander
Southern California	5	2	6 Male; 10 Female
			10 High School; 6 Junior High
Fresno	3	-	5 Suburban; 5 Urban (not inner city); 6 Inner City
Total Number of Teachers	12	4	

Making follow-up phone calls. During the three- to six-week period chosen by each teacher to be the time s/he would teach the unit and complete a portfolio, the PCP phoned the teachers to offer support, answer questions, and ascertain progress. At least one phone call was made to every teacher, and all teachers were informed that they could call collect (if necessary) if they had any questions about the portfolio activity.

Arranging for the mailing of portfolios. Each teacher was sent a Federal Express envelope in which to mail the portfolio to FWL. Whenever possible, teachers were asked to mail their portfolios in the portfolio binders given to them in August.

Recruiting and training of scorers. Three experienced high school English teachers and one college English professor who also served as coordinator of teacher preparation were recruited to score the portfolios. The three high school English teachers had previously assessed and scored the assessment center activities in the August, 1990 pilot test. For that same pilot test, the college English professor had been trained as an alternate assessor.

The two high school English teachers who had served as trainers for the August, 1990 pilot test also designed and conducted the scoring training.

Collecting evaluation feedback. FWL staff designed two evaluation feedback forms on which the teachers and scorers could give their thoughts and opinions about the portfolio assessment. The teachers mailed in their completed forms along with their portfolios. The scorers also mailed in their forms after scoring the portfolios.

Security

No attempt was made in this pilot test to verify that the completed portfolios were developed by the teachers who submitted them. In fact, as the assessment developers pointed out in their final report to the CNTP, such an attempt to ensure "originality" would be "counter to the expectations, expressed in the instructions, that classroom teachers will draw on multiple resources, [including] collaboration with experienced teachers, and may radically revise the [lesson plans] when they teach them."

To help ensure authenticity, however, the assessment developers offered several recommendations. For example, authenticity might be checked through an onsite visit by an assessor/scorer who would not be evaluating, but would observe a portfolio lesson in progress to verify that the lesson actually was taught by the beginning teacher.

Alternatively, if a follow-up feedback interview were included as part of the portfolio activity (such an interview was not part of this pilot test), certain questions could be asked to reveal the ownership of the teaching experience documented by the portfolio. Finally, because the portfolio activity, which is the fourth activity of the Secondary English Assessment, requires the teachers to provide samples of student work with their own responses to that work, the handwriting could be compared to the handwritten responses in the assessment's first activity, Activity A (Responding to Student Writing).

Another security measure to be considered if this assessment prototype is adopted is the collection and storage of the completed portfolios. For this pilot test, all teachers were given binders in which to keep their portfolios, and express mail envelopes in which to return them. Not all of the teachers used the binders, however, usually because their portfolios in the binders were too thick to fit into the return envelopes. In the future, if a portfolio assessment is adopted by the state for licensure purposes, FWL staff recommends that all portfolios be "bound" in some way (e.g., use of a binder) for easy storage purposes. Also, all portfolios would need to be retained for a minimum number of years, enough to cover the period in which teachers could appeal decisions, or to meet statutory requirements.

Assessors and Their Training

Administration of the portfolio assessment was basically a matter of providing the teachers with directions on how to complete the portfolio and then allowing the teachers to proceed with the task as best they could. As such, no assessors were needed to administer the activity. However, a FWL staff member designated as the Portfolio Contact Person (PCP), monitored the activity and, as described earlier, was available to teachers by phone to answer questions, provide information, and offer support.

Teacher and FWL Staff Perceptions of Administration

Almost 75% (11 of 15) of the teachers said that three months was sufficient time in which to construct a portfolio. Those teachers who said the time was insufficient cited "outside circumstances" as interfering with the completion of the portfolios within that time or a miscalculation on their part of the time needed for certain activities.

Approximately 75% of the teachers also said that they would have preferred another time of year in which to complete the portfolio activity. Almost all of these teachers suggested the middle of the school year or spring as better times for doing a portfolio.

Teachers offered several reasons why the start of the school year is not the best time for constructing a portfolio: the beginning of the school year is very hectic; teachers spend the first few months of the school year adjusting to their new school and new students; the first few months of the school year are spent teaching behavior expectations; and school populations are not stable the first few months of the year. One teacher in Los Angeles, for example, called the Portfolio Contact Person in November to report that a "small disaster" had happened: Due to an influx of new students to the school, there had been a massive shifting of classes and he had lost the two English classes with which he was working to construct a portfolio. As a result, he had a very difficult time collecting student evaluations at the end of the unit as is required for the portfolio.

Two teachers offered other perspectives on what is and is not a good time to construct a portfolio. One teacher suggested that it not be done around grade reports. Another would have preferred having had the chance to try out the unit first before constructing a portfolio on it.

Although the teachers were free to ask for help from other teachers while constructing their portfolios, only one teacher said he did so. This teacher received help from a CNTP mentor-lead teacher. Approximately 25% of the teachers (4 of 15) said that the phone calls to or from the Portfolio Contact Person were helpful and sufficient. One teacher was aided by a study guide for the novel being taught, and another teacher found it helpful to receive a time extension.

Almost 80% of the teachers (12 of 15) also noted that the binder provided to them for keeping their portfolio was helpful. One of the teachers who did not find it helpful explained that she didn't use it until she assembled everything at the end.

Approximately sixty percent of the teachers (9 of 15) spent one to two hours each week working on their portfolio. A little over one fourth of the teachers spent less than one hour, and the remaining teachers either spent more than two hours or the time varied depending on the week (e.g., some teachers spent more time at the end of the activity than at the beginning).

Because the teachers were informed of the criteria for scoring the portfolio, they were asked how much the criteria influenced the construction of their portfolios. Almost 75% of the teachers said the criteria had very little influence or none at all on how they constructed their portfolios. One teacher who said she spent more than two hours a week on her portfolio, explained that if she had tried to construct her portfolio so that it met the six pages of scoring criteria "it would have been overwhelming."

Based on the pilot test experience, FWL staff agree that the beginning of the school year is not the best time to administer the portfolio activity. Had the activity been administered later in the school year, it is probable that more teachers would have found the three-month time period to be sufficient. It is also probable that the quality of the portfolios might have been improved if they had been constructed later in the school year when the teachers were more familiar with their students and teaching situation (for more information on the quality of the portfolios, see the "Performance on Assessment" subsection of the section "Appropriateness for Beginning Teachers"). It seems also important that teachers constructing a portfolio have at least some access to assistance, even it is only through phone contact (e.g., a Portfolio Contact Person). Finally, although the majority of teachers did not find helpful the six pages of scoring criteria, they did find helpful the use of binders as organizational devices.

Scoring

The discussion of scoring addresses the scoring process, the scorers, and their training.

Scoring process. For the administration of this pilot test, each portfolio was scored independently by a pair of scorers. After reading a portfolio through, the scorer used the rating response form to holistically evaluate the teacher's performance in six areas. For the first two skill areas, the scorer used a three-point scale (i.e., adequate, marginal, and inadequate) to rate the teacher's performance along various criteria, and then a four-point scale (i.e., 1 to 4, with 1 = weak and 4 = strong) to give an overall rating. For the remaining skill areas, the scorer used the four-point scale for the criteria and overall ratings. Each portfolio took, on average, one hour to one hour and a half to read and score.

Characteristics of the scorers. The portfolios were scored by four veteran high school English teachers and one college English professor with experience in teacher preparation. (Two of the high school English teachers also served as trainers.) All of the scorers had experience in formal writing assessment programs involving holistic scoring of writing samples (e.g., Bay Area Writing Project) and in other language arts organizations (e.g., CLP, CATE). The college English professor currently uses portfolio assessment in one of his composition courses. All five scorers were Caucasian, three of them male, two of them female. All five also resided in northern California.

Training. As mentioned earlier, the same two teachers who trained assessors/scorers for the assessment center activities of the Secondary English Assessment also trained the scorers of the portfolio activity. The trainers met with a FWL staff person for one day in December to discuss the scoring response form for the activity and the design of the training. During this meeting, some changes were made to the format of the scoring response form, and the trainers selected the materials they would use to train the scorers (e.g., samples from portfolios completed by teachers during the development of the assessment in 1989). The trainers then conferred together at a later date and determined how they would conduct the training. The trainers decided that both the training and the scoring of the portfolios could be accomplished over a two-day period.

The training was conducted at FWL in San Francisco on Tuesday, February 5. Scoring of the portfolios began early Tuesday afternoon and continued all day Wednesday, February 6. The first part of the training consisted of a review of the portfolio activity as it is described in the teachers' orientation handbook. The scorers were then introduced to the scoring format and criteria of the portfolio scoring response form. Trainers alerted scorers to several discrepancies between what the teachers were asked to do in the handbook and the specific criteria used to evaluate the teachers' performances. For the remainder of the training, the trainers guided the scorers through each of the six sections of the scoring response form, facilitated brief discussions of the criteria, and practicing scoring samples of previous portfolio work. Some discussion followed each practice scoring session.

Perceptions of training. Although the three educators who were trained as scorers described the training they received as "very good," all three also had suggestions for improving the training. The primary suggestion was the addition of "a complete model portfolio to work with in training." Another suggestion was the development of a scoring guide which would include, for example, descriptors of adequate and inadequate portfolio components. One of the scorers also commented that the training would be improved if the "trainers were more sure of what they wanted to do."

Based on our own observations of the training and on the performance data from the activity, FWL staff agree that the training could be improved by following the above suggestions. More importantly, however, FWL staff believe that the time allotted during this pilot test for training was insufficient. Greater attention needs to be given to arriving at a consensus among scorers as to the meaning of each criterion and what types of performances reflect the different ratings for each criterion. Lacking such a consensus, the scorers in this pilot test arrived at vastly different ratings (i.e., a two point difference or more) for at least 25% of the teachers on three of the six skill areas. Training needs to consist of at least two full days, with ample opportunity for discussion of and practice using the scoring response criteria and form.

Assessment Content

The content of the portfolio assessment evaluates a teacher's skills in both general pedagogy and content pedagogy. As mentioned earlier, this evaluation focuses on three areas: a) planning and implementing a unit, b) responding to student work, and c) reflecting upon the experience in teaching the unit to gain insight for further teaching.

In the following pages, the content of the portfolio assessment is discussed along each of these dimensions:

- Congruence with the California *English/Language Arts Framework* and Handbooks;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the assessment activities;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, diverse student groups);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

Congruence with the *California English/Language Arts Framework* and Handbooks

FWL staff reviewed the portfolio assessment's components and scoring criteria to see in what ways they are congruent with California's *English-Language Arts Framework, 1987*. Because the guidelines for compiling a portfolio instruct the teacher to plan and teach a unit that is literature-based, we also looked at congruence of the assessment with California's *Handbook for Planning an Effective Literature Program, 1988*.

Table 8.2 describes the ways in which the portfolio assessment is congruent with the framework and handbook. As is evident from the descriptions, congruency of the assessment with the framework is strong. In particular, both the framework and the assessment advocate the following characteristics of good language-arts instruction:

- the integration of all elements of language--i.e., listening, speaking, reading and writing;
- a literature-based program;
- the teaching of composition, oral language, and higher-order thinking skills; and,
- the informal and formal evaluation of student work.

The portfolio assessment is somewhat congruent with the handbook as well. Both, for example, advocate the importance of choosing reading material that is suitable for students' general level of emotional and intellectual maturity. Congruency is lacking, however, with the handbook's other planning criteria which focus on depth of content and excellent language use. The portfolio assessment does not evaluate a teacher's choice of unit material in either of these two areas. The assessment also does not focus on the different stages of study--i.e., before, during, and after the reading--described by the handbook.

Extent of Coverage of California Standards for Beginning Teachers

The portfolio activity was examined by FWL staff to see how well it covered the California Beginning Teacher Standards which define levels of pedagogical competence and performance that California teacher credential candidates are expected to attain (i.e. Standards 22 to 32). The standards are reprinted below (in italics), along with an analysis of how the assessment activities correspond to each standard.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* The portfolio activity does not address this standard.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies,*

TABLE 8.2

CONGRUENCE OF THE SECONDARY ENGLISH ASSESSMENT'S PORTFOLIO ACTIVITY WITH THE ENGLISH-LANGUAGE ARTS FRAMEWORK AND THE LITERATURE HANDBOOK

English/Language Framework	Portfolio Activity
o Integrated Language Arts	-Unit should include activities that integrate the language arts.
o Literature-based Program	-Unit should incorporate work(s) of literature. -Scoring criteria evaluate teacher's skills in literature pedagogy.
o Teaching Composition	-Scoring criteria evaluate teacher's skills in composition pedagogy.
o Teaching Oral Language	-Scoring criteria evaluate teacher's skills in oral pedagogy.
o Teaching Higher-Order Thinking Skills	-Unit should include activities that encourage higher-order thinking.
o Evaluation of Student Work	-Teacher required to explain methods of evaluating students and to respond to student work. -Scoring criteria evaluate teacher's evaluation methods.
Literature Program Planning Handbook	**Portfolio Activity**
o Teaching Literature	-Teacher required to plan/teach unit that incorporates work(s) of literature.
o Criteria for Literature Selection Suitability for Students	-Teacher required to choose text appropriate to student abilities and needs.
o Depth of Content	-Not addressed.
o Language Use	-Not addressed.
o Stages of Study	-Not addressed.

activities, materials and assessment plans that are well defined and coordinated with each other. The portfolio activity requires the teacher to plan a unit of instruction with clearly-stated goals. Coverage of this standard could be improved if the activity also required the teacher to submit several lesson plans.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities, and materials that are appropriate for students with diverse needs, interests and learning styles.* The scoring criteria for the portfolio activity address this standard in several ways. For example, the teacher is expected to explain the "appropriateness of materials/activities to student needs and abilities," "include activities which integrate various modes of learning," and "incorporate multi-cultural perspectives in a variety of activities." The teacher's composition, oral, literature, and language pedagogical skills are also evaluated according to how s/he considers student needs and abilities.

Standard 25: Student Motivation, Involvement, and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* Coverage of this standard by the portfolio activity is very limited. One scoring criterion requires that the teacher "appears to use greater balance of class time for high-interest, student-centered activities." Another stipulates that the teacher responds to student writing "in a way that encourages future writing efforts." The portfolio activity does not address a teacher's ability to involve students or maintain appropriate student conduct.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* None of the portfolio's scoring criteria directly evaluates a teacher's skill in effectively communicating orally and in writing with students. One scoring criterion requires that the teacher's quality of writing in the reflective essay "models an appropriate level of competence for a professional teacher of English," but the ability to write high-quality essays does not necessarily equate with a teacher's ability to communicate effectively with students.

Standard 27: Student Diagnosis, Achievement, and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* The portfolio activity addresses this standard in several ways. First, the portfolio handbook directs the teacher to "make clear how you will evaluate student performance, and how you will make evaluation criteria clear to students." Then the scoring criteria evaluate how well the teacher does the following: "explains and justifies criteria and methods for evaluating student outcomes"; uses appropriate evaluation methods "for determining student outcomes and class progress toward goals of unit"; and makes the evaluation criteria clear to the students. The activity does not address the teacher's ability to identify students' prior attainments or if the teacher achieved his/her instructional objectives.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* While the portfolio activity requires the teacher to include activities which "encourage thoughtfulness, reflection, and higher-order thinking," none of the scoring criteria evaluate whether the teacher has improved the students' abilities in these areas.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* Two scoring criteria address this standard. The teacher is required to include activities which (1) "invite self-expression, self-discovery," and (2) "encourage development of student responsibility for own learning, and/or empower student to identify own significant questions and goals, and to evaluate own achievement in light of those goals."

Standard 30: Capacity to Teach Crossculturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socio-economic differences.* The portfolio activity does not directly address this standard. Although the activity requires that the teacher describes the student population, plans and implements a unit with multicultural perspectives, and reflects on teaching in a multicultural society, the activity does not require the teacher to identify his/her ethnicity, gender, etc. Thus, unless the teacher makes comments in the reflective essay or weekly log about the differences between the students and him/herself, it is difficult to assess whether the teacher can teach students who are of a different culture or background.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* The portfolio activity does not address this standard.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and* *develops professionally through self-assessment and collegial interactions with other members of the profession.* The portfolio activity requires the teacher to write a reflective essay and keep a weekly log. These two components enable the teacher to demonstrate self-assessment skills. The other elements of the standard are not addressed.

The extent of coverage by the portfolio activity of the California Beginning Teacher Standards is summarized in Table 8.3. The table describes how the standards are addressed (e.g., by portfolio components and/or scoring criteria), and also describes the extent of coverage provided.

Job-relatedness

The teachers unanimously agreed (15 of 15) that the skill areas assessed by the portfolio activity (e.g., planning and implementing a teaching unit; responding to student work; and reflecting upon one's experience to gain insight for further teaching) are relevant to their job of teaching. One teacher commented that the portfolio's focus on responding to student work was especially relevant.

The five English teachers who scored the portfolios also thought the portfolio activity is relevant to the job of a beginning secondary English teacher. Commented two scorers:

> *[Relevancy is] really good--it asks young teachers to demonstrate that they can prepare, present, and evaluate a unit.*
>
> *Very relevant. Choosing purposes for instruction, materials, a sequence of activities, and methods of evaluation--these all seem absolutely crucial competencies.*

Appropriateness for Beginning Teachers

The appropriateness of the portfolio activity is discussed in this section from two perspectives: (1) the perceptions of the participating teachers and scorers, and (2) the teachers' performance on the assessment.

Perceptions. When asked if they had sufficient opportunity to acquire the knowledge and skills relevant to the activity in which they participated, 14 of the 15 teachers responded positively. The one dissenting teacher, in his first year of teaching, explained

TABLE 8.3

EXTENT OF COVERAGE BY THE SECONDARY ENGLISH ASSESSMENT PORTFOLIO ACTIVITY OF THE CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Category of Scoring Criteria Addressing Standard	Extent of Coverage
22: Student Rapport and Classroom Environment	-Not Addressed	None
23: Curricular and Instructional Planning Skills	-Planning Abilities	Strong
24: Diverse and Appropriate Teaching	-Planning Abilities -Unit Design -Subject-Specific Pedagogical Abilities	Strong*
25: Student Motivation, Involvement and Conduct	-General Pedagogical Abilities -Subject-Specific Pedagogical Abilities	Limited
26: Presentation Skills	-Not Addressed	None
27: Student Diagnosis, Achievement and Evaluation	-Planning Abilities -General Pedagogical Abilities	Limited
28: Cognitive Outcomes of Teaching	-Not Addressed	None
29: Affective Outcomes of Teaching	-Unit Design	Limited
30: Capacity to Teach Crossculturally	-Not Directly Addressed	None
31: Readiness for Diverse Responsibilities	-Not Addressed	None
32: Professional Obligations	-Portfolio Presentation -Reflective Ability	Limited

*Depends on teacher providing an accurate description of his/her student population.

that he would have to have "several more years of experience in the classroom and creating my own units," before he could create a "really good portfolio."

A majority of teachers (11 of 15) also stated that they did not believe that only teachers with more than two years of experience in the classroom have the skills and abilities to pass the portfolio activity. One teacher with two and a half years experience defended a first-year teacher's ability to engage in the activity, but explained the adjustments a first-year teacher might have to make:

> *First year teachers can do it. They have to be more flexible as their time estimate is usually inaccurate (as is their estimate of ability and enthusiasm of students).*

Two other teachers thought first-year teachers could do it because the activity (i.e., planning and teaching a unit) is similar to work done for student teaching. In fact, one teacher commented,

> *I think I would have found this easier and more natural in the time immediately after I completed my credential.*

A little over 25% (4 of 15) of the teachers, however, indicated that years of teaching experience do make a difference, or, in the words of one teacher, "Veteran teachers could do a better job." A second-year teacher, for example, explained why he would not have wanted to do the portfolio in his first year of teaching:

> *I'm not extremely happy with my portfolio; however, with lesser knowledge in my first year (last year) it would have been worse.*

The five English teachers who scored the portfolios were also asked if they thought that a beginning English teacher would have had an opportunity to acquire the knowledge and skills measured by the portfolio activity. Three scorers said "yes," and two of the scorers expressed reservations. One scorer stated that teachers coming out of the state's "handful of superior schools of education" or "who work in districts with close affiliations with the writing and literature projects" would have advantages over those who did not. The other scorer thought that some school settings tended to afford teachers greater opportunities to acquire the skills and knowledge measured by the portfolio activity than others.

The scorers were also split as to when they thought the portfolio assessment should be administered in a beginning teacher's career. Three of the scorers said the second year of teaching was most appropriate, and two of the scorers selected the first year of teaching. Those in favor of a second year assessment tended to believe that beginning teachers spend much of their energy their first year just surviving:

> *A first year teacher can be so overwhelmed with trying to survive that it might be difficult to complete this activity with the detail necessary to make it an accurate, useful tool for evaluation. By the second year a teacher should have experience to draw on and increasing competence and confidence that would make this activity a more accurate indicator of skill, knowledge, and professionalism.*
>
> *The first year is too often consumed by contextual issues beyond a first year teacher's control. After surviving the first year, young teachers might have a clearer opportunity to reveal their true competence.*

One of the two scorers who advocated a first year assessment was the English professor who works with student teachers. His rationale for administering the assessment in the first year of teaching is based on the assumption that the assessment would provide feedback to the teacher:

> *If the assessment occurs after student teaching but during the first year, when the novice teacher bears full responsibility for planning, implementation, and evaluation of course work, it would not only seem most appropriate (and fair) but also most useful because it would provide feedback when the candidate needs it and is still expecting it.*

In conclusion, almost all of the teachers thought they had the opportunity to acquire the skills and knowledge measured by the portfolio activity, but some of the scorers were not so sure. Differences in teacher preparation programs, district staff development programs, and school settings were cited by scorers as possible reasons that not all beginning teachers would be able to do well on the portfolio activity. While most of the teachers also thought that the first or second year of teaching was an appropriate time to administer the assessment, some of the teachers tended to believe that more teaching experience would likely equate with higher quality portfolios. Similarly, a slight majority of the scorers named the second year of teaching as the most appropriate time for administering the assessment because the first year of teaching is usually devoted to survival, and, hence, teachers have little energy available for creating a portfolio.

Performance on assessment. FWL staff analyzed the teachers' performances on the portfolio activity to see if the beginning teachers participating in this assessment had acquired the knowledge and skill measured by this activity. Specifically, FWL staff looked at the overall ratings for each of the six evaluation categories listed on the scoring response form. Because each teacher was rated by two scorers, the ratings from both scorers were included in the analysis. Although the rating scale included four possible ratings, ranging from a high of "4" to a low of "1," the ratings were not designed with pass/fail characteristics. For our purposes, however, we interpreted the "4" and "3" ratings (4 = definite strengths in this area; 3 = some strengths in his area) as "pass" ratings, and the "2" and "1" ratings (2 = lacks strength in this area; 1 = serious weaknesses in this area) as "fail" ratings.

Table 8.4 shows the number and percentage of teachers who received "pass" ratings (i.e., "4," "3," or "3" and "4"), "fail" ratings (i.e., "1," "2," or "2" and "1") and a mixture of ratings (i.e., a combination of ratings which did not clearly indicate a "pass" or "fail" performance) for each evaluation category. As the table indicates, only one evaluation category was clearly passed by at least 50% (9 of 16) of the teachers. This category, "Reflective Ability," evaluated teachers on how well they analyzed and evaluated components of their unit, extrapolated accurate and useful information about their own teaching and about their students' abilities and needs, and reflected on the teaching of English and/or language arts in a multi-cultural society. For the remaining five categories, a little over one third to almost one half of the teachers received passing ratings.

Interestingly, the category, "Reflective Ability," was also failed by the greatest number of teachers (6), along with the category, "Subject-Specific Pedagogical Abilities." One scorer who perceived the teachers as having difficulty with writing a reflective essay described the essays he scored as often "reportial," "generically self-critical," and "sometimes self-advertising." He suggested that the teachers' difficulty in displaying a reflective writing ability "may be in part unfamiliarity with reflective and metacognitive writing and in part a problem of audience perception." He elaborated on the latter point by saying that, for the vast majority of California teachers, the kind of writing that they do in schools that is closest to the reflective writing required by the portfolio activity is usually directed to administrators. Writing for this audience, the teachers usually offer no more than "blandly sincere self-criticism," said the scorer.

TABLE 8.4

NUMBER AND PERCENT OF TEACHERS RECEIVING EACH RATING IN THE SIX EVALUATION CATEGORIES OF THE PORTFOLIO ACTIVITY

Evaluation Category	Clearly Passed ("4," "3," or "3" & "4")	Mixture (e.g., "3" & "2")	Clearly Did Not Pass ("1," "2," or "2" & "1")	Other
I. Planning Abilities	7 (43%)*	3 (19%)	5 (31%)	1 N.E. (6%)
II. Unit Design	6 (37%)	5 (31%)	4 (25%)	1 N.E. (6%)
III. Portfolio Presentation	7 (43%)	5 (31%)	4 (25%)	
IV. General Pedagogical Abilities	6 (37%)	4 (25%)	5 (31%)	1 Missing Rating (6%)
V. Subject-Specific Pedagogical Abilities	6 (37%)	2 (12%)	6 (37%)	2 Missing Ratings (12%)
VI. Reflective Ability	9 (56%)	1 (6%)	6 (37%)	

*Percent figures are rounded off to the nearest hundredth.

The table also reveals that for three of the six evaluation categories (i.e., "Unit Design," "Portfolio Presentation," and "General Pedagogical Ability"), at least 25% of the teachers received mixed ratings. That is, the teachers were given a "pass" rating by one scorer and a "fail" rating by another scorer. Such mixed ratings make it difficult to make conclusions about (1) the teachers' performances in these categories, and (2) the appropriateness of the categories for beginning teachers. (For more information about the lack of reliability among scorers, see the Format section.)

Although the assessment does not include an overall portfolio rating, and the developers did not set a pass rate for the activity (i.e., the number of categories which need to be passed in order to pass the activity), our analysis of the ratings shows that just under 50% (7 of 16) of the teachers passed four or more of the six categories. Specifically, the pass rate breakdown for the seven teachers was as follows: one teacher passed all six categories; four teachers passed five categories; and two teachers passed four categories. Of the remaining nine teachers, one passed three categories, four passed only one category, and four did not pass any category.

The pass rate breakdown described above tends to suggest that overall the portfolio assessment is a difficult one for many beginning teachers. But is this a fair conclusion? When we looked at the pairs of ratings given to the four teachers who passed only one category, for example, we discovered that three of the teachers had at least three mixed ratings and one of the three had five mixed ratings. In other words, of the four teachers who passed only one category, three of the teachers received at least one passing rating in at least four categories.

On the other hand, of the four teachers who did not pass any category, only one received a mixed rating. For the other three teachers, each pair of scorers gave the teachers' performances in each of the six categories a "1" or "2" rating.

Considering the variation in scoring, then, what can be concluded from the performance data about the appropriateness of the portfolio activity for beginning teachers? Overall, because approximately 30% or more of the teachers clearly did not pass four of the six categories, and because 25% of the teachers clearly did not pass any of the six categories, it seems safe to conclude that (1) the portfolio activity is not too easy for beginning teachers, and (2) for some beginning teachers, the portfolio activity is extremely difficult. In addition, because over one third of the teachers clearly did not pass the two evaluation categories, "Subject-Specific Pedagogical Abilities" and "Reflective Ability," these are two areas in which beginning teachers probably need more instruction, preparation, and/or experience.

Overall, the performance data tend to support the following conclusion made by one of the scorers regarding the appropriateness of the portfolio activity for beginning teachers:

> *I suspect we were testing many of our candidates on tasks that they were really not taught to do. The whole process entailed in the portfolio is quite demanding: defining the context of instruction and analytically describing the student population; choosing a focus and goals for the thematic instruction; constructing a thoughtful sequence of materials and activities; creating a method of evaluation that would be clear to the students and would actually assess whether the goals of the unit have been achieved; keeping an honest detailed log and writing a reflective essay...These are difficult tasks for a new teacher.*

Appropriateness across Contexts

In order to determine if the teachers believe the Secondary English Assessment's portfolio activity is appropriate for teachers across contexts, we specifically asked them to comment on the assessment's appropriateness for teachers of diverse student groups (e.g., different student ability levels, different ethnic groups, handicapped or limited-English students, different school/community settings). Approximately 92% (14 of 15) of the teachers responded positively to the question. We also asked the scorers to comment on this issue, as well as analyzed the performance data with respect to the performances of teachers of different grade levels and of other diverse student groups. The following two sections look at these areas in more detail.

Grade level. In this assessment pilot test, none of the teachers or scorers made any reference to the inappropriateness of the portfolio activity for teachers at different grade levels.

Analysis of the rating results, however, suggests that there may be some differences among teachers of different grade levels. For example, of the five teachers who received the lowest ratings (i.e., four teachers who did not clearly pass any of the categories; one who clearly passed only one category and had no mixed ratings), four were teachers of junior high/middle school students. Thus, of the six junior high/middle school teachers who participated in the pilot test, almost 70% clearly did not pass.

While it should be mentioned that the remaining two junior high/middle school teachers were two of the five teachers who received the highest ratings (i.e., clearly passed at least five of the six categories), the performance results discussed above still tend to indicate that the junior high/middle school teachers seem less prepared than the senior high teachers. In fact, these results are similar to the results of the Secondary English Assessment's three assessment center activities. In that pilot test, the participating junior high/middle school teachers (who were the same teachers that participated in the portfolio activity) tended to receive lower ratings than the senior high teachers in two of the three activities. Thus, in two different types of assessment pilot tests (i.e., performance-based exercises conducted at an assessment center and a portfolio assessment), the junior high/middle school teachers seemed to be less prepared than the senior high teachers. (As for possible reasons as to why this is so, we can only speculate that the most skilled secondary teachers may gravitate toward and be hired at the high school level.)

Diverse students. As reported at the beginning of this section, almost all of the teachers responded positively to the question of whether the portfolio activity is appropriate for teachers of diverse student groups. Two of the teachers, however, qualified their answers. One teacher thought the activity would be fair "only if work specific to the [diverse] student groups is requested in the portfolio." The other teacher commented,

> *Teachers of LEPs have new students thrust on them weekly (in my district) and there needs to be some design advantage to incorporate these new variables.*

The teachers' belief that the portfolio activity is appropriate for teachers of diverse student groups is strengthened by the fact that they are all teachers of diverse student groups. For example, all of the teachers who participated in the activity taught in classrooms where at least some students spoke a language besides English. In addition, almost 60% (9 of 15) of the teachers taught in classrooms where five or more languages were spoken.

However, students who speak a language other than English are not necessarily limited-English proficient (LEP). FWL staff agree with the teacher above whose comment seems to indicate that teachers of LEP students may be at a disadvantage over teachers of proficient English speakers. For example, as part of the portfolio activity, teachers are requested to "provide samples of students' work, with your written responses." Teachers with a majority of LEP students in their classes may tend to have more oral activities than written, and to respond orally as well. Although the portfolio directions allow for the submission of audio or video tapes, it seems the creation of such tapes puts an unfair burden on the teachers of these students. Audio and/or video tapes (especially the latter) cost much more money than the collection (and possible xeroxing) of students' written work, and videotaping requires equipment and expertise not readily available in all schools. Thus, although the teachers may believe that the portfolio activity is fair to teachers across contexts, it seems possible that it may be an easier activity for teachers of proficient English speaking students.

The scorers were also asked to address the issue of student diversity, but in a slightly different way. The scorers were asked to comment on how the portfolio activity addresses a beginning teacher's ability to work with diverse students. Of the five scorers, only one gave an unqualified positive response. While three of the scorers acknowledged that the activity requests that the teacher respond to student diversity, two of the scorers also pointed out a lack of evidence in many portfolios that the teacher was actually doing so.

One scorer offered a possible explanation as to why evidence of a teacher's ability to work with diverse students was sometimes lacking. Citing the directions for Part I of the portfolio which ask the teacher to "describe the student population: age range, ethnic mix, grade level, basis for placement in course, special needs or abilities," this scorer emphasized that the availability of evidence of the teacher addressing student diversity was very much dependent on "if the candidate follows directions." The scorer suggested that if the teacher provides a description of the student population, then the portfolio activity has the ability to "indicate the teacher's awareness and ability to work with diverse students." If, however, the teacher does not follow directions--or provides an inaccurate description--then the scorer's ability to evaluate how well the teacher addresses student diversity is greatly hampered or rendered impossible.

At the very least, the majority of scorers seem to believe that even if the portfolio activity is unsuccessful in addressing a teacher's ability to work with diverse students, the activity is to be commended for putting a focus on student diversity. Commented one scorer:

> *I see [the portfolio activity] as useful for focusing and emphasizing a need to reflect about true diversity. Not all candidates succeeded in addressing this issue, but most at least thought about it as a result of compiling the portfolio.*

Fairness across Groups of Teachers

When asked if they felt the assessment is fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers, all but one of the teachers and all of the scorers said "yes." (One teacher did not answer the question.)

The FWL analysis of the teachers' performance data, however, suggests a possible exception with regard to teacher preparation. Specifically, three of the sixteen teachers participating in the activity received their teacher preparation from a teacher preparation program outside California, and all three were among the five teachers who received the lowest ratings. (A fourth teacher in the group of five was prepared for teaching at a California private school.) One of these teachers commented,

> *I wish I'd gone through this when I first started teaching in California--my [out of state] experience is almost groundless here.*

Furthermore, of the three teachers who received their teacher preparation at a UC campus, two were among the five teachers with the highest ratings. Thus, although the numbers are too small to arrive at any firm conclusions, they suggest that (1) teachers who receive their teacher preparation outside the state may be less well prepared for the portfolio activity than teachers prepared by California state institutions, and (2) teachers prepared by the UC system may be more likely to achieve passing ratings on the portfolio activity than teachers from other institutions, both public and private.

Appropriateness as a Method of Assessment

In addition to evaluating the appropriateness of the portfolio activity for beginning teachers, and its appropriateness across contexts and groups of teachers, the teachers and scorers were asked to evaluate the appropriateness of the portfolio activity as a method of assessment. In particular, the teachers were asked if they thought the portfolio activity was an appropriate way of assessing their skills in the following areas: (1) planning and implementing a teaching unit, (2) responding to student work, and (3) reflecting upon their experience to gain insight for further teaching. The greatest number of teachers (92%) thought the third area was most appropriate for assessment by a portfolio. The first and second areas were judged to be appropriate for assessment by a portfolio by 79% and 52% of the teachers respectively.

387

Those teachers who did not think the portfolio method was an appropriate way of assessing their skills in planning and implementing a teaching unit usually objected to the "implementing" part. One teacher, for example, described what he saw as a weakness of the portfolio assessment with regard to implementation of a unit:

> *You are unable to see what transpired in the classroom by way of discussion, etc.*

Those teachers who did not think the portfolio method was an appropriate way of assessing their skills in responding to student work tended to have one of two objections. First were those who commented that the portfolio activity did not capture a teacher's oral responses to student work:

> *Since most all responding was done orally in class, perhaps a visitation would have been of benefit here. Students don't really read comments on papers--thus we review* ***everything*** *orally during class time.*

Second were those who thought that the samples of student work required by the portfolio (i.e., the work of one student for an entire unit and a mix of student responses for one activity) were not sufficient or somehow inappropriate for assessment of a teacher's skill in responding to student work:

> *Each assignment and student are both unique. It's difficult, I think, to really assess a teacher's abilities using such a limited scope.*
>
> *The other part of the assessment (Activity A) was better suited to that. Looking back, I see that my responses on the samples I happened to choose were not very enlightening.*
>
> *Sampling is too small. I gave a lot of feedback to some terrible papers, but I don't want to include those! I would rather show what some of the more capable students wrote.*

The latter teacher's comment seems to suggest a fear that his/her ability to respond to student work will somehow be connected with the quality of the student work submitted. In fear of being judged a terrible teacher, the teacher may choose to omit from the portfolio all student work that s/he views as terrible. Perhaps that is why the second teacher's comment advocates the assessment center activity, Activity A (Responding to Student

Writing), as a better assessment approach to this skill than the portfolio activity. In Activity A the teacher is asked to respond to two student writing samples, both of which are supplied to the teacher from unknown students. In this way, the focus is totally on how the teacher responds to particular student work, and not on the student work itself.

The scorers were asked to comment on the appropriateness of the portfolio activity as a way to assess general teaching skills and skills in teaching English classes. In general, the five scorers responded positively to both questions. Exclaimed one scorer:

> *The portfolio activity is the strongest indicator of these skills!*

Only one scorer had a dissenting opinion and this was in regard to whether the portfolio activity is an appropriate way of assessing general teaching skills. As he explained,

> ***Helping students read thematically related literary texts or to read a single text themetically is different from reading texts for information (as in science for instance). Of course, there is overlap. But generic instructional techniques tend to create awkward fits with the concrete purposes of a specific discipline.***

In conclusion, a majority of the teachers and the scorers seemed to think the portfolio activity is an appropriate method of assessment. The teachers especially favored the portfolio activity as a method of assessing reflective skills and skills in planning and implementing a teaching unit. Many of the teachers had doubts as to whether the portfolio activity is an appropriate way of assessing a teacher's skill in responding to student writing. Some thought the activity unfairly focuses on a teacher's written responses to student work, ignoring a teacher's oral responses; others thought the student writing samples required by the activity are either insufficient or inappropriate for the purpose of assessment. One scorer also expressed doubt as to whether the portfolio activity is an appropriate way of assessing general teaching skills, specifically the teaching of reading. He stated that helping students read texts in an English class, for example, is different from helping students read texts in a science class, because in the latter there is a much greater focus on strictly reading for information.

Assessment Format

This section discusses the format of the portfolio assessment from two perspectives: (1) the actual construction of the portfolio (i.e., the format of the six tasks comprising the portfolio activity), and (2) the clarity of the scoring response form and process. The format section is primarily based on the comments of the teachers and scorers, as well as the perceptions of FWL staff.

The Construction of the Portfolio

The format of the portfolio activity requires the teacher to construct a portfolio consisting of six components each of which requires the teacher to complete a different task. As described earlier, these six components are: (1) an outline of the unit plan, (2) a weekly log, (3) materials/assignments given to students, (4) samples of student work with teacher responses, (5) student evaluations, and (6) a reflective essay. The teachers were asked if they had any difficulties with any of the components, and if so, to describe the difficulties.

Overall, the component which was a source of difficulty for the greatest number of teachers was that which required the collection of student work with teacher responses. Approximately 40% (6 of 15) of the teachers had difficulty with this task, and several different reasons were offered. The most common reason cited was a problem of copying the students' work before submitting it as part of the portfolio:

> *I would forget to copy before returning materials and have to re-collect [the work].*
>
> *Copies of students' work needed to be made. This is not easy under some budgets.*
>
> *Xeroxing can be a problem. We are limited on the number of photocopies we are allowed each day (if the machine is working!).*

Other reasons given focused not on the copying but on the collecting of student work. One teacher talked of "respect" for a student's right to not submit work for the teacher's portfolio. Another teacher pointed out that "students do not always turn in their assignments", and added,

> *I felt I would have liked better samples but some of the students I had picked didn't do part of the work.*

A couple of teachers had difficulty neither with the collection or copying of student work, but instead with the requirement that the student work include their responses. One teacher explained why she did not have the time to always comment to students about their work:

> *As an English teacher I had papers daily. I struggled just keeping up with grading--forget comments.*

Another teacher claimed to provide comments, but had difficulty providing them as part of her portfolio because much of the "reading, responding and reviewing work" in her class was done orally.

A second component that was a source of difficulty for many of the teachers (5 of 15) was the task which required the development of an outline of the unit plan. Some of the teachers seemed to experience difficulty with long-range planning:

> *It's always difficult to plan long-range because circumstances (e.g., students' abilities, special schedules) may necessitate changing those plans.*
>
> *I found the outline hard because, though I do a lot of brainstorming for myself before I begin a unit, I often come up with the ideas along the way.*

Other teachers had difficulty planning for a unit with which they had little or no experience:

> *I had not read the book previously, so I didn't have an extremely clear idea about how to outline the unit in advance.*
>
> *I was prepared to teach another unit and knew what to keep. My activities were sparse in my second choice unit.*

Finally, one teacher's difficulty lay with the discrepancy between what he planned to do and what he actually did:

> *My outline was a "wishlist"--unfortunately, I didn't do everything I'd wished.*

This discrepancy turned out to be a source of difficulty for the scorers, too, several of whom noted the discrepancy in a number of portfolios.

In conclusion, of the six components required for the construction of the portfolio, the format of two proved a source of difficulty for between 30-40% of the teachers. The greatest number of teachers had difficulty with the component which required them to collect samples of student work with teacher responses. Although the teachers' difficulties were varied, FWL staff believe that several are worth paying particular attention to. First, for some teachers, collecting student work meant xeroxing student work--a task which often had to be paid for by the teacher. Second, some teachers found it difficult to collect student work because their students didn't complete the work. And third, teachers who used a lot of oral activities in their classes, and hence generally responded orally to their students, found it difficult to include such work in the portfolio. These three difficulties suggest the possibility that teachers who (a) have money or free access to a xerox machine, audio equipment, or video equipment, (b) have a majority of students who complete their assignments, and/or (c) usually use very few oral activities, may be at an advantage when completing this component of the portfolio.

The second component which caused difficulty for many of the teachers was the outline of the unit plan. Some teachers' difficulty seemed to be a lack of experience with long-range planning. The difficulty for other teachers, however, was a lack of experience with the material they were teaching. This raises the following questions: Are teachers who teach a unit with which they are already familiar (e.g., have taught the unit before) likely to be at an advantage when completing this component? If so, is there any way to ensure that teachers plan and teach a unit that is new to them? Or should the portfolio assessment only be administered during the second year of teaching after teachers have had the opportunity of planning and teaching several units during the first year?

Some teachers also had difficulty with planning a unit which they weren't able to implement as they planned. Are these teachers at a disadvantage for what might be termed "wishful" planning? One of the scorers seemed to suggest they are as he described what he considered to be a weakness of the entire portfolio assessment:

> *In most of the portfolios I read, there were discrepancies among the context, the plan, the sequence of activities, the log, the student samples and evaluations, and the reflective essay. In some cases, these discrepancies were quite minor or were explained by the candidate in the log or the reflective essay. In other cases, there was no apparent awareness of the discrepancies and no explicit comment.*

The scorer expressed his frustration to FWL staff about not knowing how to score those portfolios where the discrepancies are not explained. This is an issue that needs to be addressed if the portfolio activity is considered for further development work by the state.

Although the six components of the portfolio activity aim to reveal much about the teacher's skills in several areas (e.g., planning, general pedagogy, reflection), careful consideration needs to be given to ensuring that none of the components puts at an unfair advantage (e.g., is less expensive, is much easier) a particular type of teacher or teaching context or a particular type of portfolio. Although a teacher with better writing skills may have an easier time completing a weekly log (one of the portfolio's components), that advantage is very different from the one experienced by a teacher in a wealthy school who has ample access to xerox machines and audio/video equipment. An assessment component which assesses a teacher's writing ability is not the same as an assessment which, in effect, assesses a teacher's ability to procure equipment or copy student work. Similarly, care should be taken that a teacher who takes risks and plans a creative unit but is unable to implement that plan is not at a disadvantage compared to a teacher who plans a conservative unit (e.g., a few traditional activities) that is easily implemented.

Clarity of the Scoring System and Response Form

The scoring system and scoring response form for the portfolio activity were described briefly in the introduction and in the "Scoring" section. To recap, the scorers holistically evaluated the teacher's performance on the portfolio by using a scoring response form divided into six evaluation categories, each of which represents a different skill or competency: Planning, Unit Design, Portfolio Presentation, General Pedagogical Abilities, Subject-Specific Pedagogical Abilities, and Reflective Ability. Using either a three- or four-point rating scale, the scorers assessed the teacher's performance related to each skill/competency on two levels: (1) according to specific criteria listed for the skill/competency, and (2) an overall rating at the skill/competency level. All of the scorers were asked about their experience in rating the teachers on the portfolio activity. Their responses are discussed below.

Each of the scorers described some difficulty in scoring the portfolio activity. Interestingly, the two evaluation categories that provided the most difficulty for the scorers were also two of the categories that the smallest percentage (37%) of teachers clearly passed (see Table 8.4). These two categories were "General Pedagogical Abilities" and "Subject-Specific Pedagogical Abilities."

The "General Pedagogical Abilities" category provided some difficulty for four of the five scorers. In particular, three of the four had difficulty with the criteria which focused on evaluation methods and evaluation criteria. In the orientation handbook for the portfolio activity, the teachers are instructed, as part of their outline of the unit plan, to "make clear how you will evaluate student performance, and how you will make evaluation criteria clear to students." One scorer remarked that a description of evaluation criteria was "often missing" from the portfolios, and another scorer went further and stated, "None of the portfolios that I read described evaluation criteria." The latter scorer also added,

> *It's pretty difficult to know whether evaluation criteria are clear to students--unless that issue were addressed in the student evaluations.*

Since three of the eight criteria listed for the "General Pedagogical Abilities" category focus on evaluation methods and criteria, it is easy to see how scorers might have difficulty giving an overall rating for the section. How much weight, for example, should scorers give the three evaluation criteria in relation to the other five? And how might a teacher demonstrate in a portfolio that evaluation criteria are clear to the students? Questions such as these need to be addressed in training to make the scoring process easier.

The evaluation category, "Subject-Specific Pedagogical Abilities," proved difficult for three of the five scorers. This category requires the scorer to rate the teacher's skills according to criteria listed under four subject-specific pedagogies: composition, oral, literature, and language. Two of the scorers described the criteria as "redundant" or "repetitious," and one of them added that there were "lots of unclear items." The other scorer simply commented,

> *Oral pedagogy was almost impossible to score.*

Oral pedagogy was evaluated along two criteria, one of which reads, "Provides helpful guidance and feedback for oral performance or group work." It is not hard to see how scorers might have a difficult time with this criterion. In order to receive a rating, the teacher basically has to record his/her oral feedback to the student(s) and submit this recording as part of the portfolio. None of the teachers did this.

In the FWL staff analysis of the "Subject-Specific Pedagogical Abilities" category, we agreed with the scorer who found some of the criteria to be unclear. The most obvious example, perhaps, is a criterion used to evaluate literature pedagogy which reads, "Empowers students in reading a work to discover and respond both to the social or human significance of ideas and to the aesthetic qualities of the language." As there was no discussion of the meaning of this criterion during the training, it is very likely that each scorer interpreted the meaning of this criterion very subjectively. Indeed, when we looked at random at three pairs of scorers' ratings for this criterion for three different portfolios, we discovered that one pair of scorers gave exactly the same ratings for this criterion, but the other two differed greatly. From one pair, one scorer found "no evidence" of the criterion, while the other found evidence and rated it a "2" (i.e., serious weakness in this area). The other pair differed even more, giving the respective ratings of a "2" and a "4" (i.e., definite strength in this area). Although these are only two examples of how one criterion can be interpreted differently by different scorers, FWL staff believe that at least half of the twelve criteria listed under the "Subject-Specific Pedagogical Abilities" section are similarly susceptible to subjective interpretation by the scorers.

We also looked at the issue of redundancy or repetition and again concurred with the scorers who stated that such redundancy or repetition exists. For example, one criterion used to evaluate literature pedagogy reads, "Chooses text(s) appropriate to student abilities and needs, and goals of unit, within constraints of school and classroom context." This criterion seems redundant considering that, in the scoring response form's first category, "Planning Abilities," the teacher is evaluated according to how well s/he "explains appropriateness of materials/activities to classroom context (e.g., administrative and material constraints)," "explains appropriateness of materials/activities to student needs and abilities," and "relates choice of materials/activities to goals of unit."

Similarly, a criterion used to evaluate oral pedagogy reads, "Provides a range of speaking opportunities appropriate to student needs and abilities, goals of unit and classroom context." As was pointed out above, the teacher's choice of activities, including those which afford speaking opportunities, is evaluated in the first category according to the appropriateness to student needs, etc. Although the criterion used to evaluate oral pedagogy focuses on the range of appropriate speaking opportunities versus just their appropriateness, there is still a sense of repetition in the language used. In fact, because there are numerous criteria on the scoring response form which are similar in language, but different in meaning, it seems very important that the differences between them be emphasized during training (which they were not for this pilot test) or that they be rewritten to more clearly highlight the differences.

In addition to describing scoring difficulties, the scorers made suggestions as to how they would like to see the scoring process and response form revised. Some of the suggested revisions were directly related to the difficulties described above. Two of the scorers, for example, advocated "tightening" the scoring response form by collapsing or eliminating some of the criteria, thus producing a scoring form that is "clear, focused, and streamlined."

Another suggestion for revision focused on the order of the six evaluation categories as they are presented on the scoring response form. One scorer questioned "the logic" of placing the evaluation category, "Portfolio Presentation," in the middle of the scoring response form. He pointed out that this category is distinctly different from the other categories because it does not directly relate to teaching skills. As he explained,

> *Preparing a portfolio is not a teaching skill, but an activity that assesses and enhances those skills.*

He recommended moving the evaluation category, "Portfolio Presentation," from the middle of the scoring response form to either the beginning or the end.

Consideration might also be given to eliminating this category altogether since six of its nine criteria do not measure teaching competencies, but rather the ability to follow directions (e.g., "Student evaluations include full set," "Student samples include evidence of teacher's response strategies"). Criteria which focus on how well the teacher organizes and/or presents different components of the portfolio do not seem to equate with criteria which evaluate the teacher's teaching skills. Perhaps in replacement of this evaluation category, there could be a checklist which serves to ascertain if all of the components of the portfolio are complete and present. If the teacher's portfolio is substantially deficient in this area, then consideration should be given to returning the portfolio to the teacher without evaluation.

Three other suggestions for revision were to (1) include an overall rating for the portfolio activity, (2) include criteria that address internal consistency among the components of the activity, and (3) conduct a task analysis for each section of the portfolio to make certain that there are criteria keyed to each task. FWL staff agree that an overall rating for the portfolio activity should be considered, and, as we discussed earlier, that the issue of internal consistency among the components (e.g., Is that which is described in the unit plan evidenced in the other components?) needs to be addressed.

We also agree with the third recommendation that a task analysis be conducted for each component of the portfolio to ascertain the match between the work the teacher is required to do and the scoring criteria that assess that work. For example, one of the six components of the portfolio activity, referred to as Student Evaluations, instructs the teacher "to collect written evaluations from students, either of the entire teaching unit, or of one major activity." Although this component comprises one sixth of the portfolio activity, only 2 of the 46 scoring criteria listed on the scoring response form correspond to this component. Moreover, one of the two criteria simply ascertains whether "the student evaluations include a full set (most of class)." FWL staff question how a criterion of this kind evaluates a teacher's professional competency. The second criterion, "Student evaluations provide useful feedback to teacher for evaluating unit or specific activity," is closer to a more appropriate assessment criterion, but does not seem to go far enough. That is, instead of putting a focus on the kind of information provided by the student evaluations, the focus should be on what the teacher does with or how the teacher reacts to the information.

In conclusion, based on the feedback from the scorers and on FWL staff analysis, the scoring response form for the portfolio activity needs extensive revision. Consideration should be given to revising the form so that all scoring criteria are easy to understand, distinctly different (i.e., not repetitious), and measure true teaching competencies. Revisions should also be made to ensure a tight match between the criteria and the portfolio components. Most likely, the scoring process as a whole would benefit if the number of scoring criteria were substantially reduced. This would facilitate training, reducing the amount of time needed to develop distinctions between each level of competency for each criterion. Finally, the scoring process should be revised to include an overall rating for the portfolio activity.

Cost Analysis

We can use the experience of administering the portfolio activity of the Secondary English Assessment as a basis of estimating the costs of implementing a portfolio assessment. To review, the piloted portfolio activity consists of six types of documentation of a three- to six-week unit and is completed over a three-month period by the teacher. Each portfolio is reviewed and scored independently by a pair of scorers. A six-part scoring system is used, where each part represents a different skill or teaching competency. Each skill is evaluated according to a list of specific criteria as well as an overall rating for that skill. All of the ratings are made in a holistic manner and are not interdependent. Below, we outline the general assumptions and basis for estimating the costs of administering and scoring portfolios, and training scorers. It is important to view these as only general, incomplete estimates. For example, a more complete and specific total estimate would include the costs of developing assessments and training materials, selecting and recruiting trainers and scorers, and training trainers...

Also presented below is a brief discussion of the development and pilot testing costs of this activity.

Administration and Scoring Cost Estimates

Administering the portfolio assessment is basically a matter of providing teachers with directions on how to complete the portfolio and then allowing the teachers to proceed with the task as best they can. Thus, no assessors are required for administering this assessment.[1]

Scoring, on the other hand, requires trained raters who are knowledgeable in the content and criteria for the assessment, and about a variety of teaching contexts. Experience with the pilot test suggests that a half day of scorer training is insufficient, and we recommend at least two full days of training. This extended training would include a complete sample portfolio as a training model, and ample opportunity for discussion of, and practice using, the scoring response criteria and form.

In our pilot test, after scorers had some practice and experience scoring, each was able to read and score one portfolio in an average of one to one and a half hours. The actual pilot test involved five scorers evaluating 15 portfolios (a pair of scorers for each portfolio) over a period of one and a half days after a half day of training. An analysis of the pair ratings for each portfolio indicated that consensus among scorers was not high (i.e. there was a two-point difference or more for at last 25% of the teachers on three of the six skill areas rated). Though the extended training included in this cost estimate should improve the reliability of scoring, it is still likely that a third scorer will be necessary on some occasions to resolve discrepancies. In addition, someone is needed to periodically monitor scorers to ensure that they are applying scoring criteria correctly. Past experience with other assessments we pilot tested, indicates that a trainer can serve both of these purposes, or, in future assessments, we assume an experienced scorer could fulfill the role of "lead" scorer.

---

[1] Experience administering the portfolio activity of the Secondary English Assessment pilot test suggests that someone in a district needs to be trained in how to construct and score a portfolio to: 1) answer teacher questions and provide assistance in constructing a portfolio, and 2) monitor teacher progress and completion of a portfolio. If mentor teachers or other district personnel are trained, then these tasks would not represent additional administration costs. Otherwise, these tasks are an additional cost of administering the portfolio assessment.

Based on our experience, we assume that a portfolio assessment can be scored on a regional basis. We estimate that approximately five days per scorer would be required for each of ten scorers to score 20 portfolios. We assume that one of the ten scorers would fulfill the role of lead scorer.[2] According to this logic, ten scorers should be able to score 200 portfolios, or the portfolios of 100 teachers (i.e., a pair of scorers independently scores each portfolio) over the five-day period. Using a rate of $160 per day for each scorer and $210 for a lead scorer yields an estimate of $86.50/assessment for a portfolio to be scored by two scorers.[3]

We estimate that training for this assessment should be at least two days. Two hours for preparation and finalizing logistics is needed for each scorer. If we assume that each scorer would score 40 portfolios each year for three years, we could distribute the costs for training a scorer over 60 assessments. After the first year, we estimate a day of training and preparation would be needed for each scorer each year that would include any changes in the assessment criteria or procedures. Reimbursing scorers for four and a quarter days of training at $160/day or $20/hour would cost $680. Distributing the $680 over the 60 assessment adds approximately $11/assessment for training.

Other costs would include those associated with duplication of materials, postage and travel where needed. Based on our experience from the various pilot tests, a figure of $30 per assessment assumes only minimal travel costs for a regional assessment. This is the same estimate we used in the cost analyses of other assessment instruments that we pilot tested.

---

[2] We assume that the lead scorer would probably score less than 20 portfolios given the additional responsibilities of resolving discrepancies and monitoring the scoring. Based on past experience with the pilot tests, we assume other scorers will become more proficient at scoring and make up the difference.

[3] This figure is calculated based on nine scorers at $160/day for five days which would cost $7,200 and one lead scorer at $210 per day for five days which comes to $1,050, for a total of $8,250. Distributing the costs over 100 portfolio assessments yields an estimate of $82.50/assessment.

In summation, our estimates for administering and scoring a portfolio assessment are as follows:

Scorer Costs:	$82.50/assessment
Training Costs:	$11/assessment
Other Costs:	$30/assessment
Total Admin/Scoring	$123.50/assessment

Development and Pilot Testing Costs

As mentioned earlier, the portfolio activity is one of four activities which were developed as part of the Secondary English Assessment. In Chapter 7, we described the other three activities (i.e., the assessment center activities) of the assessment, and stated that the development costs for all four activities totalled $84,415. These costs are broken out by cost categories in Table 8.5. The total figure represents all the expenses the assessment developer incurred in delivering drafts of the four activities to the CTC and SDE. In developing these activities, the developer was not initiating a completely new development effort; instead, the developer was building on prior work with these and other assessment activities. Hence, any future development costs are likely to be similar to those incurred for this pilot test.

The costs incurred for pilot testing the portfolio activity were approximately $ 57,384. These costs are also broken out by cost categories in Table 8.5 .

Technical Quality

This section discusses the technical issues (development, reliability, and validity) related to the portfolio activity associated with the Secondary English Assessment.

TABLE 8.5

PILOT TEST COSTS FOR THE PORTFOLIO ACTIVITY

Cost Categories	Pilot Testing
Staff-Salaries & Benefits	$16,662
Consultants (Teachers, assessors, and other consultants)	20,198
Travel (Consultants and staff)	6,467
Other Direct Costs (phone, shipping, duplication)	2,036
Total Direct Costs	$45,363
Indirect Costs	12,021
Total Costs	$57,384

Development

The portfolio activity was part of the Secondary English Assessment whose development was discussed in Chapter 7. It was administered and scored at a later date. The portfolio activity constituted a different assessment approach than the performance-based assessment center activities of the Secondary English Assessment. Therefore, it is being analyzed separately, but its developmental history is the same as was described for the Secondary English Assessment. Minor improvements were made in the portfolio handbook which contained directions for completing the activity.

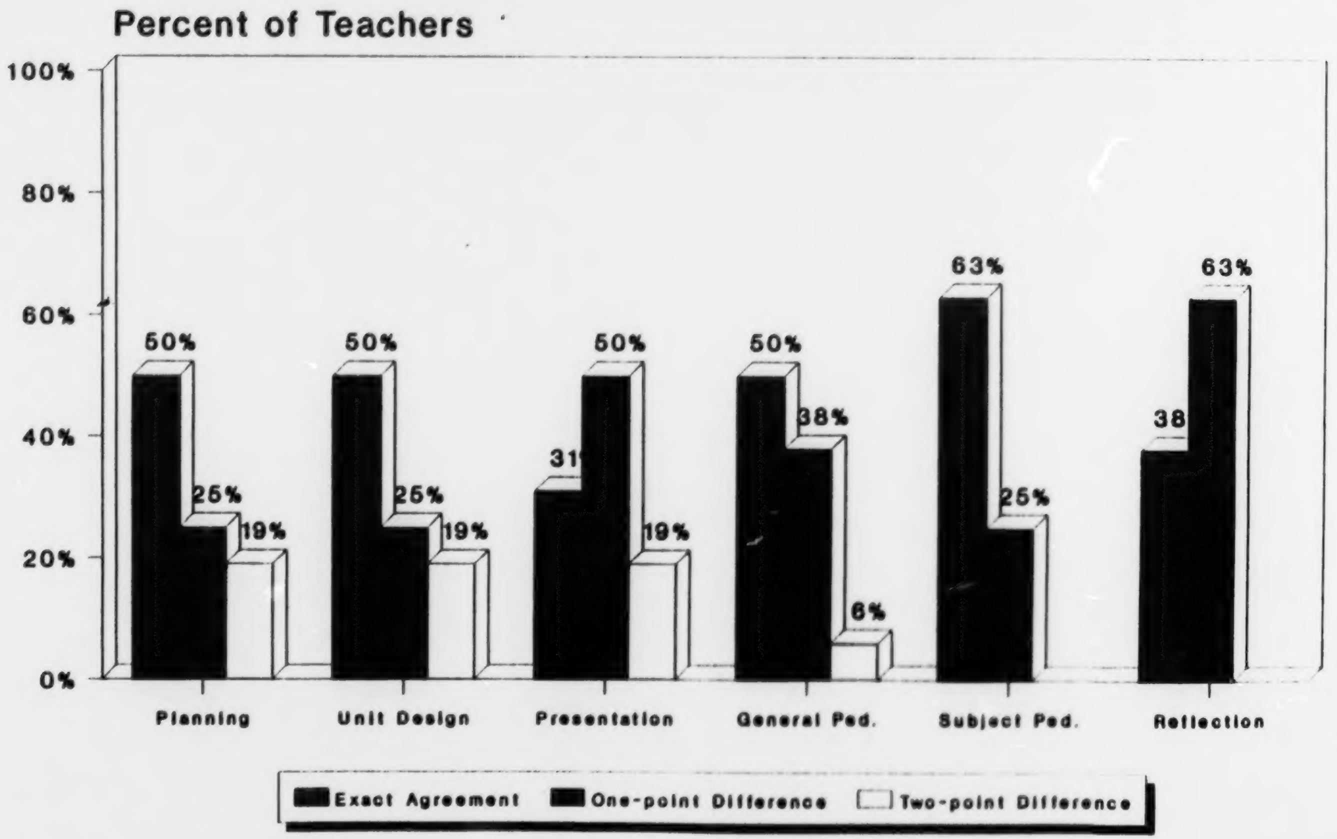
Reliability

The following analyses were performed for the portfolio activity using the pilot test data of 16 teachers. Inter-rater agreements were examined to assess the degree to which assessors were able to consistently judge candidates using the English-Language Arts Assessment scoring protocols. Internal consistency estimates were generated to assess the degree to which the variables or factors within each of the parts would form a measure and the degree to which the different parts related to each other and might form an overall assessment of a candidate's performance on the portfolio activity.

Inter-rater agreements. The first measure of agreements among judges was obtained by comparing the number and percent of ratings in which raters gave identical or different ratings. Figure 8.1 presents the percent of exact agreements for the six parts of the Portfolio Activity. They range from a low of 31 percent for Portfolio Presentation to a high of 63 percent for Reflective Ability. Most other pairs of ratings differed by one point. Raters differed by two points for three teachers each for Planning Activities, Unit Design, and Portfolio Presentation, and for one teacher in General Pedagogical Skills. In addition, two teachers received a rating from one rater of "N.E.", meaning that the rater believed that there was not enough information to rate the portfolio in this category. This level of agreement on the Portfolio Activity suggests that a moderate degree of agreement has been achieved through the training and scoring associated with the pilot test. As mentioned earlier, more specification of criteria coupled with more lengthy training is needed to raise the level of inter-rater agreement.

Internal consistency of the tasks and assessment. Coefficient Alpha reliability estimates were calculated for the six parts within the Portfolio Activity by using the individual ratings on items within each part. The reliabilities for the overall Portfolio Activity and its parts are listed in Table 8.6. These estimates indicate a very high degree of

FIGURE 8.1
Percent Agreement Rates for the Portfolio Activity



*Percentages do not total 100% because in five cases one rater only believed there was insufficient information to that ma omp ratin pos

TABLE 8.6

INTERNAL CONSISTENCY OF THE PORTFOLIO ACTIVITY AND ITS PARTS

Part	Reliability
Planning Abilities	.93
Unit Design	.89
Portfolio Presentation	.86
General Pedagogical Abilities	.86
Subject-Specific Pedagogical Abilities	.87
Reflective Ability	.87
Summed Ratings Across All Parts	.90

internal consistency within the parts and the summed ratings of the parts for the entire Portfolio Activity. A likely factor that contributes to this high internal consistency is that raters rated the overall parts of the portfolio in a more holistic manner and did not make explicit, independent ratings on the subparts. This suggests that either holistic ratings will be more efficient than using subpart-based ratings or that greater training and emphasis in rating subparts will be needed to capture the potential analytic information for subparts.

Intercorrelations among activities. Correlations among the six parts of the Portfolio Activity were calculated for the teacher candidates who received clear ratings on specific pairs of parts. The number of teachers ranged from 11 to 15 for each pair of parts. The correlations appear in Table 8.7.

Despite the relatively small N, many of the correlations were statistically significant. Part I, Planning Activities, was the only part that did not yield significant intercorrelations with other parts. For the 16 candidates in the pilot test the internal consistency across parts was .90. This provides tentative evidence that the parts as presently defined tend to measure the same attributes of the teacher candidates' performance, implying that a "total" rating across the parts could be used to make a judgement on the teachers' skills on the overall portfolio.

Validity of Agreement Through Group Comparisons

Differences in performances were examined for minority-non-minority, women-men, high school-middle school, and suburban-urban-inner city teachers. It was felt that this could provide at least preliminary glimpses of the assessments difficulty for different groups. Some of these analyses that compare different groups have been discussed in earlier sections. The pilot test sample size and design were not constructed to provide information sufficient to provide stable estimates comparing differences among these groups. For example, some subgroups have as few as two teachers in them. Nevertheless, an examination of differences among groups provides some initial insights into the validity of this assessment. Table 8.8 contains a summary of the trends for the pilot sample of 16 teacher candidates. Appendix F provides the means, standard deviations and numbers of candidates from which these summaries were constructed. A plus (+) simply indicates that the mean or average for the first group was greater than that for second group. For example, the pluses under the Female-Male column indicate that for four of six variables, the average score of females was greater than that of the males.

TABLE 8.7

INTERCORRELATIONS AMONG PARTS OF THE PORTFOLIO ACTIVITY

	I	II	III	IV	V	VI
Part I: Planning Activities	--	.15	.44	.33	.24	.60*
Part II: Unit Design		--	.85***	.77**	.70	.49
Part III: Portfolio Presentation			--	.91***	.84***	.62*
Part IV: General Pedagogical Abilities				--	.93***	.78**
Part V: Subject-specific Pedagogical Abilities					--	.87***
Part VI: Reflective Ability						--

* $p < .05$
** $p < .01$
*** $p < .001$

TABLE 8.8

TRENDS OF MEAN DIFFERENCES BETWEEN CANDIDATES WITH DIFFERENT CHARACTERISTICS FOR PARTS OF PORTFOLIO ACTIVITY

Part	Gender: Male/ Female	Level of Teaching: H.S./ M.S./Jr. High	Teaching Location: Suburban, Rural Inner City, Urban	Ethnicity: Non-Minority, Minority
I: Planning	–	+	–	+
II: Unit Design	–	+	–	+
III: Portfolio Presentation	+	+	–	+
IV: General Pedagogical Abilities	+	+	–	–
V: Subject-specific Pedagogical Abilities	+	+	+	+
VI: Reflective Ability	+	+	+	+
SUMMARY	4/6	6/6	2/6	5/6

*Entries reflect the direction of the mean differences for the different candidates. For example, in Part I: Planning Activity, the average or mean of female teachers in the pilot test was less than the males'. These do not generally represent statistically significant differences and due to small Ns no tests of significance were calculated.

Trends in the table provide some preliminary evidence that on this assessment and pilot test sample:

- females' averages were greater than males on 4/6 evaluation categories;
- high school teachers' averages were greater than junior/middle school teachers' on 6/6 evaluation categories;
- teachers in suburban settings had averages that were greater than those in urban/inner city settings on 2/6 evaluation categories; and
- non-minority teachers' averages were greater than minority teachers' on 5/6 evaluation categories.

If these trends were to hold for larger, more representative samples, some of these trends provide less than encouraging findings. For example, if non-minority teachers continue to outperform minority teachers, it suggests that the instrument may be biased toward non-minority teachers (or, alternatively, minority teachers need to strengthen their performance in this area relative to non-minority teachers).

Content validity. The content validity of this assessment rests largely in the role that teachers and English educators had in the development, and the analyses of the match of the assessment to the model curriculum and teaching standards which supplies evidence that the assessments content validity with current and emerging content. These have been described earlier and implications for further development are described in the following section.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Secondary English Assessment's portfolio activity. The section presents information in the areas of administration, content, and format, and concludes with a brief summary.

Administration of Assessment

For the portfolio assessment, the teacher plans and conducts a three- to six-week teaching unit, compiling a portfolio of six components to document the activities within the unit. Thus, the time allotted for administration of this type of assessment is much longer than for other assessments, requiring up to several months to allow the teacher the flexibility to plan and implement the teaching unit within the constraints of his/her teaching context.

Based on our experience, in order to ensure a smooth implementation of the portfolio activity, we offer the following **recommendations** :

- administer the activity at a time other than the beginning of the school year because beginning teachers generally spend the start of the school year adjusting to their new school and new students, teaching behavior expectations, and coping with constantly shifting class populations;

- allow a minimum of three months for teachers to complete the portfolio;

- make available to the teachers some form of assistance (e.g., Portfolio Contact Person accessible by phone) during the time they complete the portfolio;

- provide binders which the teachers can use to help organize the portfolio; and,

- develop procedures to collect and store the completed portfolios.

In this pilot test, the teachers received an orientation handbook which described in detail the required components of the portfolio activity and the scoring criteria used to evaluate the activity. Although the orientation materials that provided instructions to the teacher on how to complete the portfolio were perceived as helpful by all the teachers, the six pages of scoring criteria proved to be too long and detailed to be of any help. In any future administrations of the portfolio activity, it would probably be beneficial to reduce the number of scoring criteria so as to make it easier for teachers to refer to them when constructing the portfolio.

In addition to the construction of the portfolios, the administration of the portfolio activity included the training of scorers and the subsequent scoring of the portfolio activity. Both the training and scoring were conducted over a two-day period, with the majority of time devoted to scoring. Based on the scorers' comments, our observation of the training, and on performance data from the portfolio activity, we believe training could be improved by following these **recommendations**:

- recruit scorers who are experienced English teachers, who have had experience with holistic scoring, and who are knowledgeable about a variety of teaching contexts;
- develop a scoring guide which includes performance markers for different ratings;
- use or develop a complete sample portfolio to serve as a training model;
- extend training to at least two full days; and,
- provide ample opportunity for discussion of and practice using the scoring response criteria and form.

Following the above suggestions should greatly facilitate the administration of the portfolio assessment activity.

Assessment Content

The portfolio assessment activity focuses on a teacher's skills in three areas: (1) planning and implementing a unit, (2) responding to student work, and (3) reflecting upon the experience in teaching the unit to gain insight for further teaching. Based on the observations of FWL staff, as well as information from scorers, teachers, and the assessment documentation (i.e., the scoring response forms), the following conclusions are offered about the content of the portfolio activity.

- Congruence of the portfolio activity with the *English-Language Arts Framework* is strong. In particular, both advocate an English-Language Arts program that integrates all elements of language, is based on literature, teaches composition, oral language, and higher-order thinking skills, and includes an informal and formal evaluation of student work.
- Coverage by the portfolio activity of the eleven California Standards for Beginning Teachers is relatively weak. Although the activity does a good job of addressing the standards which focus on curricular and instructional planning skills and diverse and appropriate teaching, coverage of the other nine standards can best be described as limited or nonexistent.

- The teachers and scorers unanimously agreed that the skill areas assessed by the portfolio activity are relevant to their job of teaching.

- Although almost all of the teachers thought they had the opportunity to acquire the skills and knowledge measured by the portfolio activity, the performance data tended to indicate otherwise. Looking at the ratings for the teachers' performance in each of the six evaluation categories, our analysis shows that less than 50% (7 of 16) of the teachers clearly passed four or more of the six categories.

- Because 25% of the teachers clearly did not pass any of the six categories, it seems safe to conclude that for some beginning teachers, the portfolio activity, as presently designed, is extremely difficult.

- Based on the fact that over one third of the teachers clearly did not pass the two evaluation categories, "Subject-Specific Pedagogical Abilities," and "Reflective Ability," these are two areas in which beginning teachers may need more instruction, preparation, and/or experience in order to do well.

- Teachers and scorers thought the portfolio activity is fair to teachers of different grade levels. However, analysis of the performance data seems to suggest that junior high/middle school teachers may be less well prepared than senior high teachers to do well on the activity.

- Almost all of the teachers thought the portfolio assessment is appropriate for teachers of diverse student groups, a perception that is strengthened by the fact that all are teachers of diverse student groups. However, because one of the portfolio's six components require samples of students' work, some including the teacher's written responses, it is possible that teachers of a predominantly limited-English proficient class would have a more difficult time with this assessment since the majority of their student work and responses to such work is likely to be oral.

- When asked to comment on how well the portfolio activity addresses a teacher's ability to work with diverse students, only one of the scorers gave an unqualified positive response. Other scorers pointed out that the ability of the activity to evaluate a teacher in this area depends wholly on the teacher providing an accurate description of his/her student population.

- With regard to the question of the assessment's fairness across groups of teachers (e.g., different ethnic groups, different language groups), both the teachers and the scorers responded positively. The performance data suggests, however, that teachers who receive their teacher preparation outside the state may be less well prepared for the portfolio activity than teachers prepared by California state institutions.

- When asked to comment on the appropriateness of the portfolio activity as a method of assessment, a large majority of the teachers found the portfolio activity to be an appropriate method of assessing reflective skills and skills in planning and implementing a teaching unit. Far fewer teachers (52%) thought the portfolio activity is an appropriate way to assess a teacher's skills in responding to student work. Some thought the activity unfairly focuses on a teacher's written response to student work, ignoring a teacher's oral responses; others thought the samples of student work required by the activity are either insufficient or inappropriate for the purpose of assessment.

Assessment Format

The format of the portfolio assessment requires the teacher to perform six distinct tasks, all within the framework of implementing a teaching a unit. Four of the tasks involve written responses from the teacher (i.e., outline of plan, weekly log, written responses to student work, and reflective essay), and three require the teacher to choose, collect, and/or organize material such as student assignments, student work samples and student evaluations of the unit. (There is overlap on one of the tasks.) In addition to the construction of the portfolio, our discussion of the portfolio format included an analysis of the assessment's scoring process and scoring response form. Based on comments by teachers, scorers, and FWL staff, we offer the following conclusions and recommendations regarding the construction of the portfolio and the format of the scoring process and response form:

Construction of the Portfolio. Two of the portfolio tasks proved to be a source of difficulty for 30-40% of the teachers.

- The task which was a source of difficulty for the greatest number of teachers (approximately 40%) was that which required the collection of student work with teacher responses. Although the teachers' reasons varied, the most common difficulty cited was a problem of copying the students' work before submitting it as part of the portfolio (i.e., teachers either had limited access to a copying machine or had to pay to have student work copied). Teachers also found it difficult to collect samples of work from students who didn't do or complete the work, and to collect samples of student work during predominantly oral activities.

- The second task which caused the most difficulty for many of the teachers (30%) was the outline of the unit plan. Some teachers' difficulty was a lack of experience with long-range planning. A difficulty for other teachers was a lack of experience with the material they were teaching.

Based on the above findings, we offer the following **recommendation**:

- Review all tasks of the portfolio to ensure that none puts at an unfair advantage a particular type of teacher or teaching context. As currently designed, some of the portfolio tasks may be easier for teachers who (a) can easily afford or have free access to a copy machine, audio equipment and video equipment, (b) have a majority of students who complete their work, (c) usually use very few oral activities, and (d) teach a unit with which they are already familiar (e.g., have taught the unit before).

Scoring Process and Response Form. Based on feedback from the scorers and on FWL staff analysis, the scoring response form for the portfolio activity needs extensive revision. In particular, revisions to the scoring criteria should be made following these **recommendations**:

- Eliminate/revise those criteria that are strongly susceptible to subjective interpretation (e.g., avoid phrases such as "respond...to the significance of ideas and to the aesthetic qualities of the language").

- Rewrite those criteria that are similar in language but different in meaning so as to more clearly highlight the differences.

- Include only those criteria that measure true teaching competencies. Eliminate all criteria that solely assess how well the teacher has followed directions (e.g., six of the nine criteria in the evaluation category, "Portfolio Presentation").

- Eliminate all criteria in which there is not a tight match between the criteria and the portfolio tasks.

- Consider reducing the number of criteria by which the teacher is evaluated.

In addition to revising the scoring criteria, we offer the following **recommendations**:

- Consider developing within the scoring process the means to address the issue of internal consistency among the portfolio tasks (e.g., Is there evidence that what has been planned has been implemented? If not, what does the scorer do?).

- Because the majority of criteria listed under the evaluation category, Portfolio Presentation, do not directly relate to teaching skills, consider eliminating this entire category.

- Consider developing a new category that does not purport to evaluate a teacher's skills, but rather serves as a checklist to ascertain if the teacher followed directions when compiling the portfolio (e.g., Are all portfolio components complete and present?). If a substantial portion of the portfolio is incomplete or missing, consider developing a system whereby that portfolio is returned to the teacher without evaluation.

- Include an overall rating for the portfolio activity.

Summary

As currently designed, the portfolio activity of the Secondary English Assessment constitutes an innovative method of assessing a teacher's reflective skills and skills in planning and implementing a teaching unit. Although designed to also assess a teacher's skills in responding to student work, the portfolio activity seems less aptly suited for this task, in part because of the various problems many of the teachers experienced collecting student work for the portfolio. The portfolio activity has the potential for providing rich and bountiful information about a teacher's skills in the English/Language Arts area, but substantial and extensive revisions would have to be made to the assessment's scoring system before this potential could ever be realized.

CHAPTER 9

SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES

The Semi-Structured Interview in Secondary Social Studies (SSI-SSS), developed by the Connecticut State Department of Education, assesses the depth, breadth, and accuracy of the knowledge of a beginning secondary social studies teacher in three areas: curriculum content, content pedagogy, and knowledge of students. First, a teacher performs a task resembling part of their teaching responsibilities, and then responds to a series of structured questions addressing their rationale for the choices they made. The pilot tested interview focused on the teaching of the Pre-Civil War Era in U.S. History, using the following five tasks:

- **Unit Planning**. A teacher arranges a set of cards representing topics (i.e., major events or societal features) relating to the Pre-Civil War Era, and then responds to questions concerning the ordering of the topics for effective teaching and learning.

- **Use of Documentary Materials**. A teacher reviews a variety of documentary materials (e.g., documents, charts, and pictures) that might be used to teach a specific topic, and designs a lesson around a subset of those materials. The teacher then responds to questions concerning the possible uses of the documents in teaching U.S. history.

- **Historical Interpretation**. A teacher reads several historical interpretations of the causes of the Civil War, and develops approaches that might be used to help students understand the different interpretations presented as well as the concept of historical interpretation in general.

- **Alternative Pedagogical Approaches**. A teacher examines five different activities to acquaint students with the pre-Civil War era, describes how each might be used in an average ability social studies class, and discusses the advantages and disadvantages of each approach.

- **Evaluating Student Learning**. A teacher develops an essay question to evaluate student learning on the topic of slavery, and answers a series of questions addressing learning objectives, characteristics of a high quality student response, anticipation of problems students might have, and modification of the essay

> question for students of varying ability in social studies. The teacher then turns to two essays from his/her actual class, one representing an average response and one a weak response. After describing the classroom context, the teacher discusses the choice of essay question, explains his/her comments made on the two papers and describes alternative methods of evaluating whether students have achieved the same learning goals.

The SSI-SSS is modeled after the Semi-Structured Interview in Secondary Mathematics (SSI-SM). The SSI-SM was piloted in California during the first year of pilot testing, and is analyzed in the *Assessment Component of the California New Teacher Project: Year One Report.* One purpose of developing a second interview was to explore the instrument's ability to be generalized across subject areas.

Since the scoring system has not been developed yet, the interview is videotaped to allow scoring at a later date. The scoring process is holistic, with trained scorers comparing the teacher responses to sample performances of "marker" teachers representing different degrees of breadth, depth, and appropriateness.

Administration of Assessment

This section provides an overview of the assessment administration, and a discussion of the logistics of administration, security issues, and assessors and their training: it concludes with a summary of teacher and assessor impressions of administration.

Overview

The Semi-Structured Interview in Secondary Social Studies (SSI-SSS) was administered at one site in the Bay Area in December, 1990. As seen in Table 9.1, a total of 17 teachers credentialed in secondary social studies participated, split roughly evenly between males and females. Fourteen of the teachers were white; two were black and one was Hispanic. Most (13 of 17) taught at the high school level; roughly one-third had previously taught a unit on the pre-Civil War era, the topic on which the assessment focussed. Twelve of the teachers were in their first or second year of teaching; the remaining five were in their third or fourth year.

Some difficulty in recruiting beginning teachers with experience teaching U.S. History was experienced. Several teachers told us that in their schools, experienced teachers chose to teach U.S. history, while a typical assignment for beginning teachers was

TABLE 9.1

PILOT TEST PARTICIPANTS

SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES

(Number of Teachers = 17)

Descriptive Characteristics of Participants	Distribution of Participants
Gender	
Male	9
Female	8
Ethnicity	
Black	2
Hispanic	1
White, Non-Hispanic	14
Grade Level	
Middle School/Junior High School	4
High School	13
Previous Experience in Teaching Unit on Pre-Civil War Era	
Yes	6
No	10
No Response	1
Number of Years Taught, Including Present Year	
One	4
Two	8
Three	4
Four	1

World History. They added that they hoped to eventually teach U.S. History after they acquired more seniority, indicating a belief that teaching U.S. History was a perquisite reserved to reward experienced teachers. And indeed, in calling most districts from Fremont to Sacramento, we found that beginning teachers teaching U.S. History classes were rare.

Each interview lasted approximately four hours. A modified assessment center approach was used to administer the interview. Instead of having teachers rotate to different stations for different tasks, each teacher was administered all tasks by a single team of two interviewers. Generally, while one interviewer administered a task, the other operated the video camera and periodically monitored it. The roles were reversed after every one to two tasks. In seven of the seventeen interviews, one task was omitted because the interview was badly behind schedule. For the most part, the two teams of interviewers completed four interviews per day; two teachers were interviewed in the morning and two in the afternoon. The interviews were conducted over a six-day period, with a half-day break in the middle.

Logistics

Logistical activities for this assessment included identifying teacher samples and preparing orientation materials.

Identifying teacher samples. The identification of teacher samples proved to be more difficult for secondary social studies teachers than for teachers of other subject matters, as beginning social studies teachers seem to be at lower densities within a concise geographic area than teachers in math, science, or English. Even when beginning teachers who taught social studies were located, they often were ineligible for the pilot test, as they were credentialed in another subject area. Given the short time frame available for recruitment, we quickly decided to concentrate our search in one major metropolitan area, recruiting 17 Bay Area teachers teaching at school locations that ranged from Sacramento in the north to Milpitas in the south. Originally, we had planned to confine the sample to teachers with experience in teaching U.S. History, the focal topic of the assessment. However, as mentioned earlier, we quickly discovered that a more typical assignment for new social studies teachers was World History. In addition, several new social studies teachers contacted taught economics and political science classes, but no history classes. Therefore, although we attempted to recruit teachers who had taught U.S. history either as part of their student teaching or regular teaching assignment, at least six of the seventeen teachers had no previous experience in teaching U.S. history.

Since the assessment occurred relatively early in the school year (i.e., first week of December), we included teachers in their first, second, and third year of teaching. This was in contrast to most previous assessments occurring later in the school year, where third-year teachers were not considered to be "beginning" teachers.

Orientation to the assessment. When asked to participate, teachers were told orally about the videotaped interview format, and the need to bring to the interview two samples of corrected student essays where students were asked to write on a topic for a minimum of thirty minutes, either in class or outside of class. One of the student essays was to represent a weak response, the other, an average response. Written materials included a list of the tasks on which they would be interviewed, together with a sample question for each task to indicate the type of questions they would be asked.

In response to the request to provide student essays, many teachers anticipated difficulty. Many teachers, especially middle school teachers, explained in response to the initial oral request that they did not regularly ask their students to write essays. At least two teachers taught E.S.L. classes, and they reported that an essay was clearly beyond their students' abilities. In response to these concerns, teachers were asked to provide samples of student writing on a topic, whether or not they were essays, and to allow students to spend approximately thirty minutes on this task.

Security

Standard procedures for safeguarding confidential materials were followed during the assessment. It is not clear whether disclosure of the exact questions would constitute a security breach, if the topic were not revealed. The questions mainly prompt the teachers to explain the reasoning underlying their choices of how to perform the tasks with respect to the specific topic and materials presented. There are few "generic" answers that could be memorized; therefore, it is a researchable question whether both the general task descriptions and the specific questions could be disclosed without impacting the reliability of the assessment.

Assessors and their Training

The four assessors, two males and two females, conducting the interviews were all experienced Connecticut social studies teachers; three taught at the high school level, and one at the middle/junior high school level. The three days of training included instruction on interviewing (e.g., use of probes), some practice interviews of beginning teachers, and a subsequent session critiquing the teachers' interviewing performances. Each interview was conducted by a male/female team of interviewers.

All four assessors rated their training as "very good," the highest rating possible. All four cited the practice sessions as particularly valuable, and one also identified the discussion of when and where to probe as very useful. Suggestions for improvements from the assessors included additional practice interviews, and "talking to teachers of different backgrounds (ethnic, racial) and in different settings (urban, rural, suburban.)" One identified problem area was the use of probes. (Probing was also identified as a problem area in the semi-structured interviews on mathematics.) Despite the major emphasis on probes during training and the provision of two-and-a-half pages of guidelines specifically addressing probes, inconsistencies in the content and frequency of probes among the social studies assessors suggest that this facet of conducting the interview continues to be problematic.

All assessors agreed that interviewers need a basic to very good knowledge of the subject area and extensive teaching experience in order to accurately monitor teacher responses. One interviewer mentioned the importance of experience in teaching social studies in establishing the legitimacy of the assessment:

> *...this is seen as a legitimate process because those doing the interviewing are also teachers. If this was to be done by others, however capable and conscientious, a critical element would be lost and the entire process tarnished.*

Throughout the week of pilot testing, the Connecticut assessors expressed amazement at the California teaching context and admiration for the willingness of the California beginning teachers to teach under such conditions. Evidently, class size in Connecticut in the assessors' districts is 20 students, compared to 30 students or more in the classes of the California beginning teachers. In addition, California teachers have multicultural classes, while the classes of the Connecticut assessors are primarily white. The Connecticut assessors' classes are also part of a system which tracks students by ability, whereas that was true of only one California teacher's school, and he reported that his school was switching to heterogeneous grouping in the next school year. Lacking a control group of California assessors, it was not possible to determine any influence that these contextual differences had on interviewing practices; none was readily evident in the interviews observed. Implications of these contextual differences for the content of the SSI-

SSS and the impact on teacher responses will be discussed in a later section on Appropriateness across Contexts.

Teacher and Assessor Impressions of Administration

Teachers overwhelmingly (81% or 13 of 16) described the administrative arrangements as "reasonable". Those suggesting improvements referred either to their fatigue or to the geographic location.

In responding to a similar question, one assessor pronounced the arrangements reasonable and two did not (the remaining one skipped that page of the form). The interviewers' schedule began with preparation for the teachers about 7:30 a.m. and concluded with a debriefing at approximately 6:00 p.m.. This was thought to be demanding by three of the four interviewers, with one describing the process as "mentally exhausting" and another noting that sometimes it was difficult to listen carefully for 4-5 continuous hours.

Scoring

Scoring Process

The scoring was intended to be parallel to the scoring developed for the semi-structured interview in mathematics. The plan was to concentrate initially on scoring the first two tasks: **Unit Planning** and **Use of Documentary Materials**. The process consisted of the following steps: (1) review the videotapes of the social studies interviews using the scoring indicators used in the math interview, (2) adjust the indicators as necessary, (4) select "marker" performances for each task, using veteran social studies teachers to validate the selections, and (5) select experienced social studies teachers and train them to score the assessment, double scoring each tape as a reliability check, and (6) analyze the scoring results. Unfortunately, due to unanticipated budget problems in the Connecticut Department of Education, the scoring process was put on hold early in the first step. The remainder of this section addresses conclusions drawn from the scoring work that was completed.

The unit of scoring for the Connecticut-developed semi-structured interviews is at an indicator level, with indicators being specified areas of knowledge or abilities related to teaching. The indicators for the social studies interview were intended to be as parallel as possible to those for the math interview. The math indicators fall into three areas --

content/curriculum, content pedagogy, and knowledge of students -- with two indicators per area. The math indicators are:

Curriculum Content

- Understands principles, skills, and concepts of mathematics.

- Understands interrelationships among topics and organizes content on the basis of relationships.

Content Pedagogy

- Understands effective practices, successful approaches, and potential problems associated with mathematics instruction.

- Understands effective instructional practices that facilitate learning and are independent of the subject area.

Knowledge of Students

- Justifies instructional practices and approaches on the basis of student background and interests.

- Justifies instructional practices and approaches on the basis of student ability.

Early analysis of the videotapes by the scoring designers suggests that the content pedagogy and knowledge of students indicators transfer well from math to social studies for the tasks viewed (**Unit Planning** and **Use of Documentary Materials**). However, the two content/curriculum domain indicators used in mathematics are difficult to distinguish in social studies, as the significant principles, skills, and concepts of the social studies involve relationships -- across disciplines, to prior knowledge, and to general themes. Therefore, developers proposed to combine these two indicators previously used for the mathematics interview into one indicator for the curriculum/content domain in social studies. The developers were also disappointed in the amount and quality of information elicited by two sets of questions: (1) the questions that appear in each task that focus on how student backgrounds and abilities affect instructional decisions, and (2) the section of questions that focus on the two "essays" brought to the assessment by the teachers.

Unfortunately, since scoring was not performed, the impact of the contextual differences between Connecticut and California on the evaluation of teacher responses was not able to be determined. It was quite clear from the responses of the beginning teachers, however, that the terms "average ability" and "high ability" have quite different meanings in the two states, and these terms would need to be more clearly defined if they continue to be used in the future. In addition, a California assessment needs to more carefully define the classroom about which the teacher is speaking, either by having the teacher describe the classroom they have in mind or by more clearly specifying such factors as: whether the classroom is heterogeneous or homogeneous with respect to reading and writing ability; the proportion of limited English-proficient students; and the general ability level of students.

Assessment Content

The focal topic of the assessment was the pre-Civil War era in U. S. history. Within this topic, the interview focussed on the breadth, depth and appropriateness of teacher decisions made in relation to the prescribed tasks and the rationales underlying those decisions.

This section evaluates the content of the SSI-SSS along the following dimensions:

- Congruence with the History-Social Science Framework;
- Extent of coverage of the California Standards for Beginning Teachers;
- Job-relatedness of the instrument;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, subject areas);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

The first two dimensions, curriculum congruence and standards coverage, are based on FWL staff's analysis of the documents involved. Discussions of the remaining dimensions are based on the perspective of the participating teachers, assessors, and FWL staff as reflected in feedback forms, informal discussions with teachers and assessors, and analysis based on observation of the assessment administration.

Congruence with California Model Curriculum Guides and Frameworks

The California Department of Education periodically produces subject-specific documents, curriculum guides and frameworks. These documents serve as public statements describing the curriculum which experts in the relevant content and pedagogy believe is most appropriate for California students at different grade levels. The most recent document pertaining to social studies is the *History-Social Science Framework* (California State Department of Education, 1987). Since this framework is intended to guide instruction in California public schools, the content of the SSI-SSS, although developed for use in Connecticut, is compared with the framework in order to ascertain its congruence with the framework and to identify any needed revisions for use in California. These comparisons are summarized in Table 9.2.

While specific content is discussed with respect to specific grade levels, there are three goals which are intended to be represented across grade levels. The three goals are: (1) knowledge and cultural understanding, (2) democratic understanding and civic values, and (3) skills attainment and social participation. Each goal is accompanied by specific curriculum subgoals or strands. (The word strand is chosen to symbolize the interdependency and interweaving of the various subgoals such that none are intended to stand alone.) However, for purposes of comparison, each stand will be discussed separately.

The first goal, the goal of knowledge and cultural understanding, has six strands: historical literacy, ethical literacy, cultural literacy, geographic literacy, economic literacy, and sociopolitical literacy. Aspects of historical literacy mentioned in the *History-Social Science Curriculum Framework* are exhibited in several tasks. Historical empathy, for example, is a focus of the **Use of Documentary Materials** task. A sense of time and chronology underlies the tasks of **Unit Planning** and **Historical Interpretation**. Understanding of cause and effect, continuity and change, and history as common memory with political implications, are essential for performing the **Historical Interpretation** task.

The strand of ethical literacy is exhibited by the choice of the pre-Civil War era as a topic and the attendant emphasis on slavery in many of the tasks. The **Use of Documentary Materials** task asks teachers to explain how they would use documents to enhance students' understanding of slavery. The choice of documents for this particular task forces attention to the ethical issues underlying slavery; the other tasks provide opportunities for teachers to address ethical issues, but do not require them to do so.

TABLE 9.2

COVERAGE OF THE CALIFORNIA HISTORY/SOCIAL SCIENCE FRAMEWORK BY THE SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES

Content	Method of Coverage	Extent of Coverage
Goal of Knowledge and Cultural Understanding		
-Historical Literacy	-Use of Documentary Materials -Unit Planning -Historical Interpretation	Full
-Ethical Literacy	-Use of Documentary Materials	Partial
-Cultural Literacy	-Unit Planning	Limited
-Economic Literacy	-Unit Planning	Limited
-Sociopolitical Literacy	-Unit Planning -Historical Interpretation	Limited
Goal of Democratic Understanding and Civic Values		
-National Identity	-Topic of Task -Unit Planning -Use of Documents	Full
-Constitutional Heritage	-None	None
-Civic Values, Rights and Responsibilities	-None	None
Goal of Skills Attainment and Social Participation		
-Participation Skills	-Alternative Pedagogical Approaches	Limited
-Critical Thinking Skills	-Evaluating Student Learning	Limited
-Basic Study Skills	-Use of Documents -Evaluating Student Learning	Full

The strand of cultural literacy is minimally represented in the **Alternative Pedagogical Approaches** task, where one of the approaches includes music which is contemporary to the era studied. Opportunities to display geographic literacy and discuss how it might be promoted in students appear in the **Unit Planning** task when ways of sequencing and relating the topics of Northern industry and cotton culture are discussed.

Arguments illustrating economic literacy are represented in one of the analyses presented in the **Historical Interpretations** task, which presents a theory of the causes of the Civil War. Also, in the **Unit Planning** task, two of the topics, Northern industry and cotton culture, provide opportunities to discuss economic issues. However, the curriculum framework discusses economic literacy in terms of more basic concepts (e.g., economics as the management of scarce resources, different types of economic systems) which are not required to be addressed by any of the tasks. Similarly, for sociopolitical literacy, both the **Unit Planning** and **Historical Interpretation** tasks provide opportunities to discuss the relationship between the political system and the social system of the pre-Civil War United States, but do not require teachers to do so. Some of the topics in **Unit Planning** (e.g., Kansas Nebraska Act, Compromise of 1850) illustrate the intersection of the social and political systems.

The second goal advocated by the framework is that of democratic understanding and civic values. It consists of three curriculum strands: (1) national identity, (2) constitutional heritage, and (3) civic values, rights and responsibilities. With respect to national identity, the choice of the pre-Civil War era as the focal period of history provides opportunities to convey the pluralistic and multicultural nature of American society, to understand the American creed as an ideology extolling equality and freedom, and to recognize the status of minorities and women in different times in American history. Several topics in the **Unit Planning** task (e.g., Slavery, Dred Scott Case) would be difficult to explain without reference to one or more of these aspects to explain their significance. In addition, the documents used in the **Use of Documents** task focus on the institution of slavery and how it affected both white and black Southerners.

The strand of constitutional heritage is not directly addressed by any of the tasks, although the discussion of slavery and the laws resulting from compromises between Northern and Southern politicians represented in the **Unit Planning** and **Use of Documents** tasks would provide opportunities to raise constitutional issues with students, and at least one teacher proposed to do so. The strand of civic values, rights and responsibilities was not addressed by any of the assessment tasks.

The third goal represented in the framework is that of skills attainment and social participation. It has three curriculum strands: (1) participation skills, (2) critical thinking skills, and (3) basic study skills. One of the alternative teaching methods in **Alternative Pedagogical Approaches** is cooperative learning, where a teacher could be expected to address the development of participation skills as a potential strength of the teaching method. In **Evaluating Student Learning**, the teachers are explicitly asked to design an essay question requiring the students to demonstrate "one or more of the skills of analysis, synthesis and evaluation." Although the scoring system has not been developed at this time, it is likely that teachers designing lessons or describing the use of materials which are limited to recall and recognition and which fail to teach students to define and clarify problems, judge information related to a problem, and solve problems and draw conclusions would not score highly. The strand of basic study skills, as defined in the framework, is at the heart of the **Use of Documentary Materials** and **Evaluating Student Learning** tasks. The **Use of Documentary Materials** task asks teachers to describe how they would guide students to use three of the six skills listed: (1) to acquire information by reading primary and secondary source materials, (2) to locate, select and organize information from written sources, and (3) to read and interpret tables and political cartoons. The **Evaluating Student Learning** task focuses on how teachers evaluate whether students can organize and express their ideas clearly in writing.

In summary, although the content of the SSI-SSS was not developed with reference to the *California History-Social Science Framework*, it addresses nearly all of the strands identified under the three goals in at least a limited way, and three of the eleven strands receive extensive coverage.

Extent of Coverage of California Standards for Beginning Teachers

The California Standards for Beginning teachers are criteria for teacher competence and performance which the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. Listed below are brief italicized descriptions of Standards 22 through 32 of the *Standards of Program Quality and Effectiveness, Factors to Consider and Preconditions in the Evaluation of Professional Teacher Preparation Programs for Multiple and Single Subject Credentials* which pertain to expectations of student competencies to be attained prior to graduation from teacher preparation programs. (The remaining standards address programmatic requirements.) To evaluate this assessment instrument and to make inferences about the assessment approach which it represents in terms of the appropriateness for use with California secondary social science teachers, the stimulus materials were compared with the 11 California Beginning

Teacher Standards. These comparisons are summarized in Table 9.3. Each standard will be discussed separately.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* None of the tasks address this standard. It is conceivable that questions might be added that address aspects of this domain that are especially important to social studies instruction, such as fostering mutual respect among students while they discuss a controversial topic.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* The **Unit Planning** task require a teacher to sequence a set of topics and provide a rationale for that sequence. The **Use of Documentary Materials** task require the development of a lesson plan using a subset of the documents provided.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* Every task contains one or more questions about how a teacher would adapt their strategies if the general ability level of students changed and one or more questions about relating the strategy chosen to student backgrounds and interests.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* None of the tasks specifically addresses this standard, although most provide but do not require opportunities for teachers to talk about motivation with respect to student interest and involvement. However, the proposed scoring indicators include two that address the teacher's knowledge of student background and interests and the ability to plan activities for students of varying achievement levels. A scorer's judgement of the effectiveness of the activities described by teachers in motivating and involving students would affect the rating on the subject-specific content pedagogy indicator.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* None of the tasks

TABLE 9.3

EXTENT OF COVERAGE BY THE SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Tasks or Scoring Criteria Addressing Standards	Extent of Coverage
22: Student Rapport and Classroom Environment	-None	None
23: Curricular and Instructional Planning Skills	-Unit Planning -Use of Documentary Materials	Partial
24: Diverse and Appropriate Teaching	-Unit Planning -Use of Documentary Materials -Historical Interpretation -Alternative Pedagogical Approaches	Full
25: Student Motivation, Involvement and Conduct	-Knowledge of Students' Scoring Domain	Partial
26: Presentation Skills	-Curriculum Content Scoring Domain	Limited
27: Student Diagnosis, Achievement and Evaluation	-Use of Documentary Materials -Evaluating Student Learning	Partial
28: Cognitive Outcomes of Teaching	-Use of Documentary Materials -Historical Interpretation	Partial
29: Affective Outcomes of Teaching	-Content Pedagogy Scoring Domain -Knowledge of Students' Scoring Domain	Limited
30: Capacity to Teach Crossculturally	-None	None
31: Readiness for Diverse Responsibilities	-Unit Planning -Use of Documentary Materials -Historical Interpretation -Alternative Pedagogical Approaches -Evaluating Student Learning	Partial
32: Professional Obligations	-None	None

directly address this standard. Teachers' use of terminology and explanation of key concepts would give some indication of the effectiveness of their subject-specific communications, and would be scored under the Curriculum Content domain. However, their explanations to a peer during the interview would only be a proxy for their explanations to students in the classroom, and the strength of the relationship of interview behavior to classroom behavior would need to be investigated before using the proxy measure with any confidence.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* Prior content knowledge of students is the subject of one question in the **Use of Documentary Materials** task, and evaluating student achievement is the focus of the **Evaluating Student Learning** task. However, only one method of evaluating student learning, the essay, is fully addressed, and many teachers felt that this method had severe disadvantages for assessing the learning of their students because of its dependence on writing abilities which many of their students had yet to develop. The assessment developer recognizes this problem and plans to rethink this section.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* Evaluating information and thinking analytically is the focus of two of the tasks, **Use of Documentary Materials** and **Historical Interpretation**. Reaching sound conclusions is not directly addressed, but might be incorporated into the redeveloped task of **Evaluating Student Learning**.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* No task directly addresses this standard. To some extent, the fostering of positive student attitudes toward the subjects learned overlaps with student motivation (see Standard 25). The fostering of positive student attitudes toward themselves would most likely be scored as evidence under Knowledge of Students indicators, and fostering the ability of students to become independent learners would be scored as evidence under one of the Content Pedagogy indicators.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender,* *linguistic and socioeconomic differences.* No task addresses this standard, although many of the teachers gave multiple examples of how they would teach culturally diverse students, students of a particular gender, or limited English-proficient students in the lessons that they designed. Additional questions asking how teachers might deal with hypothetical situations portraying different types of students could be added to the interview to address this standard.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* Teaching students of diverse abilities is addressed by questions in all of the tasks. Teaching students of diverse ages is not addressed, as teachers are allowed to respond as if they were teaching students in either junior high/middle school or high school. To assess this standard, teachers might be asked how their strategy might differ if they were teaching the other group of students.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interaction with other members of the profession.* This standard is not addressed by any of the tasks, and a new task on interaction with parents or colleagues would need to be designed to address it.

The SSI-SSS focuses on content knowledge, content pedagogy, and knowledge of students. Of the eleven relevant teaching standards used by the Commission on Teacher Credentialing to evaluate teacher preparation programs, the SSI-SSS best assesses the standard on diverse and appropriate teaching. It partially assesses four other standards which focus on content pedagogy. It provides limited information with respect to two other standards focussing on communication and affect, and provides no information with respect to three other standards with varying foci.

Job-Relatedness

Teachers and assessors each completed questionnaires which asked their perceptions of various aspects of the assessment instrument. The discussions in the next five sections draw heavily from these questionnaires, beginning with questions addressing the job-relatedness of the instrument.

Teacher perceptions. As was true of the other performance assessments piloted, the SSI-SSS was designed to be more reflective of the tasks that teachers do than traditional multiple-choice tests. This was reflected in teacher evaluations of the assessment. Teachers overwhelmingly (94% or 15 out of 16) agreed that the tasks chosen for this assessment were relevant to their job of teaching secondary social studies. Some sample comments include:

> *Very relevant, diverse.*
>
> *Absolutely! I felt this gave me the opportunity to demonstrate my competency and reveal my weaknesses.*

The one teacher who disagreed taught middle school LEP students, and felt the tasks were appropriate "only insofar as I can modify them for my students, all of whom are LEP (Limited English-proficient)."

A few teachers had suggestions for making the assessment more job-related. One teacher, for example, suggested replacing the task on historical interpretation with one focussing on motivating students or making history personally relevant to students' experiences. Two other teachers noted that while the tasks were relevant, social studies covered more than just history classes. One of the comments, though approving the choice of topic, illustrates the problem in choosing a topic to assess competency in a field that covers a number of disciplines:

> *Even though I don't teach U.S. History, all Social Studies teachers should have a background in U.S. curriculum. I'm curious about assessing social science content in things like psychology, economics, sociology, government, political science, etc. A social studies credential legally allows you to teach in all those areas, even if you have NO course work -- only a passing grade on one NTE in [Social Studies] -- I see a major problem there. Depending on scheduling difficulties in a given school a person with a SS credential and NO BACKGROUND can be "forced" to teach in areas they know NOTHING about -- certainly not the best situation for students.*

These difficulties are not unique to social studies, and the same problem was identified with respect to an assessment in science.

Assessor perceptions. All of the assessors agreed that the tasks were relevant for new social studies teachers, describing them as "general areas which every social studies teacher should know" or as requiring a "new teacher to examine and explain his/her knowledge and practices." One assessor believed that the tasks were "[g]ood -- especially if one has had experience with the topic," but left unaddressed the appropriateness of the tasks for teachers who had not previously taught the topic.

Appropriateness for Beginning Teachers

Participating teachers and assessors were asked their perceptions of the appropriateness of the tasks for beginning teachers. Teachers were specifically asked whether they felt that they had sufficient opportunities to "acquire the knowledge and abilities needed to respond in a reasonable manner to the assessment questions." Sixty-nine percent (11 of 16) of the teachers agreed, some of whom made comments such as the following:

> *The assessment focused on tasks that I have encountered in my actual class situation.*

Some teachers, however, qualified their affirmative response, citing the dependence of their answer on the degree of their experience, especially with the focal topic:

> *But I'm not sure I would have after student teaching only. It would have been better if the area was one I had taught -- world instead of U.S. history.*
>
> *If I had not recently taught pre-Civil War I would probably have complained that as a beginning teacher, I had no time to prepare.*

When the teachers were asked if they had found any of the tasks to be too difficult, 63 % (10 of 16) of the teachers responded, "Yes." Half of the sixteen teachers named **Historical Interpretation** as too difficult, and three named **Unit Planning**. Teachers identifying **Historical Interpretation** as difficult gave the following explanations:

> *Historiography is something I am not extremely familiar or comfortable with. We received no training about it [in my teacher preparation program.]*
>
> *Only because the questions asked during the interview were complicated and hard to follow. It also isn't something I have done or even thought about with ninth graders.*

> *Not enough information was given about the authors [of the three interpretations provided]. One needed to have more than basic knowledge of the subject to evaluate the task to design a lesson.*
>
> *Particularly since I'm not a history teacher.*

Teachers identifying **Unit Planning** as difficult explained that they had not taught the specific topics chosen or that they felt constrained using the given topic, but did not specify the reason(s).

Of the three assessors responding to a similar question about opportunities for beginning teachers to acquire the skills being assessed, two believed that most of the teachers demonstrated the background to respond to the questions, although one of the two identified knowledge of students as a potential problem area. The third assessor said, "A new teacher may not have had the opportunity to use all of his/her skills or knowledge and so what he/she could draw from personal experience might be limited."

FWL staff observing the assessment administration noted that despite explicit instructions specifying the type of students for whom the teacher was to plan activities, the beginning teachers made repeated references to their own students, who were, for the most part, different from the descriptions provided. This was not unique to the SSI-SSS, but was typical of every assessment piloted. FWL staff suspect that beginning teachers do this because of their limited familiarity with teaching, so that they cannot draw upon their experience to make a variety of distinctions between types of students, particularly for topics which they have not taught. For beginning teachers, assessments should either focus on the type of students taught by the teacher OR repeatedly remind the teacher about the type of students on which to focus. The latter approach complicates scoring, as teachers' differential familiarity with the specified type of students will likely affect the depth and breadth of the responses.

Social studies appears to be a particularly challenging subject for assessment designers. **Mathematics**, the subject of the previous semi-structured interview developed, is hierarchically structured so that there are basic concepts with which it can be assumed that competent mathematics teachers should be intimately familiar. However, in history, the ability to display familiarity with basic concepts (assuming they could be identified) is dependent on knowledge of the specific events in the focal historical period. Only one-third of the participating teachers had taught a unit on the pre-Civil War era, the topic which was the focus of the assessment. Although almost all of the teachers did well in explaining more general concepts such as "state's rights" and "cotton culture", several had difficulty chronologically placing specific events such as the Kansas-Nebraska Act or the Dred Scott decision or in recalling details of the events. The participating teachers believed that their experience teaching the topic affected their responses, an assertion that can eventually be empirically tested. However, if the teachers are correct in their belief, this varying familiarity would need to be taken into account in scoring, which the current scoring system does not do.

Appropriateness across Contexts

The appropriateness of the assessment across contexts was examined along two dimensions: grade level and diversity of students.

Grade level. Teachers overwhelmingly (88% or 14 of 16) agreed that the assessment is appropriate across grade levels. The two teachers who disagreed cited a single task, **Historical Interpretation**, as inappropriate for junior high or middle school teachers.

Diverse students. Teachers also overwhelmingly (88% or 14 of 16) agreed that the assessment is appropriate for teachers of diverse student groups (e.g., different student ability levels, different ethnic groups, handicapped or limited English students, different school/community settings). A sample of comments supporting this opinion include:

> *Exceptionally appropriate, because it addresses issues not addressed in current tests.*
>
> *Most teachers have diverse groups, especially in ability level. But I think the teachers in special education or ESL would need different activities.*

Some teachers, though agreeing that the assessment was appropriate for teachers of diverse student groups, also thought that the assessment could be improved in this area. For example, one teacher commented:

> *But the assessment could/should be geared more specifically toward addressing these areas (more than just the general references on the card). Maybe a little too open-ended in this area -- teacher didn't have to touch on these aspects.*

Both of the two teachers who did not believe the assessment was appropriate for teachers of diverse student groups were teachers of classes composed of students limited in English proficiency. The teachers' explanations of why they found the assessment inappropriate were as follows:

> *A lot of modification is required to use these materials with limited English students and for students of different activity levels. I think your assessment tool is rather narrow.*
>
> *I had a particularly difficult time gauging what was meant by the "average" student. As an LEP teacher in the social studies, teaching the content is* *very* *different than in "regular" classrooms. For example, I could not give a class a 30 minute essay question as you asked me to write in the fifth part of the study. These students have special needs. I think you could have a fine teacher who would not do well on the assessment simply because he focuses on different language objectives.*

When the Connecticut assessors were asked if they believed that the assessment addressed the ability of a new teacher to work with diverse student groups, two of those who responded agreed that it did. The third assessor, however, reported mixed feelings: "The assessment deals with highly capable and less capable students but with homogeneous grouping. Many schools are moving away from this. If [diverse student groups] refers to ethnicity, the assessment doesn't address ESL situations."

Assessors were also asked about the suitability of the assessment for new teachers in different school and community settings. Again, two of the three who responded that it was suitable, with one noting that "in responding to a number of questions, new teachers have the opportunity to discuss school and community settings and how they might influence their practices." The third assessor, however, had some reservations about the assessment's suitability for beginning teachers in different school settings: "[There are] some problems here... for example, question 12 in task 2 based on the statement, 'Blacks in the ante-bellum South were better off under slavery.' I just wouldn't envision the question being asked in an urban school with 80 % black enrollment."

FWL staff found that the assessment, perhaps because it was developed for the Connecticut context, seemed to make assumptions about teaching contexts that were not warranted in the State of California. First, the questions about students of varying abilities focus on classrooms of students of homogeneous ability, whereas informal conversations with the participating teachers suggested that the vast majority of them taught in classrooms which were heterogeneous with respect to student ability. Second, the **Evaluating Student Learning** task assumes that part of the responsibilities of a social studies teacher is to develop students' abilities to write a cohesive essay on a topic. According to informal conversations with the participating teachers, this task was beyond the ability of many of their students, especially those for whom English was a second language.

These assumptions about teaching conditions were built into the assessment structure and questions. This mismatch between the assumptions and the California teaching context resulted in teachers being asked about teaching under conditions with which they have little or no familiarity. The most clear example was the **Evaluation of Student Learning** task where it was assumed that at least some students in the teachers' classes would be accomplished essay writers; this did not prove to be a valid assumption for the sample of teachers participating in the assessment. In addition, the assessment questions for this task seemed to implicitly confound social studies ability with writing ability. While the fostering of writing and reasoning skills is a legitimate part of social studies pedagogy, FWL staff believe that any California social studies task focusing on the evaluation of student learning must either accommodate a great range of student writing skills or allow for alternative methods of student assessment.

Given the differences in context between California and Connecticut, it is not clear that the terms such as "average social studies ability" and "highly capable students" have the same meaning to teachers in both contexts. Although the ESL teachers clearly perceived that their students were not "average," the other teachers all made repeated references to their own students when asked to focus on a classroom of students of "average" ability. Because of the general lack of details in the responses to the question of how the lesson would be modified for "highly capable students", it was difficult to identify specific characteristics of the students that the teacher had in mind when replying to that question. A more precise definition of the ability level intended appears to be needed, perhaps with reference to concrete examples of student ability (e.g., highly capable students are characterized by the ability to read and comprehend complex materials and to write a coherent essay) or perhaps with reference to normed national test scores.

Many of the participating teachers spoke quite eloquently before and after the assessment about how they taught social studies to their own particular students. In the interview, however, despite expansion of the questions explicitly aimed at eliciting knowledge of students, the few questions that specifically asked about how student backgrounds, interests, and ability levels affected the choice of activities did not always reflect the same depth of knowledge displayed at other times. One of the problems was the common task direction to plan for a specific type of students, generally a classroom of students of average social studies ability. Teachers generally ignored this direction and focussed instead on their own students while completing the task. This was evident in their repeated references to "my students" in their responses. This was true not only in the social studies and mathematics interviews, but in teacher comments on virtually every assessment instrument that asked them to respond to a set of hypothetical students. Beginning teachers have experience with a limited range of students, generally confined to those encountered in student teaching and their first year(s) of teaching. Perhaps it is not surprising that in assessments deliberately postponed until the first years of teaching on the grounds that there are skills which only develop with experience, beginning teachers do, in fact, draw heavily on their teaching experience, which is by definition limited. Because beginning teachers can often talk in great detail about decisions with respect to their own students and cannot talk about other types of students in the same depth, FWL staff believe that the interview questions should ask teachers to describe their own students and plan with respect to them. After this information has been collected, teachers can then be asked how their decisions would differ with other types of students.

Fairness across Groups of Teachers

Teachers were asked if they believe the assessment to be fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers. They overwhelmingly (88% or 14 of 16) agreed that it is fair. Some supporting comments are as follows:

> *Regardless of our backgrounds, we use the same content (in general) and similar techniques -- or at least are taught/told to!*
>
> *I am not aware of issues in this assessment which would affect teachers differently on the basis of their ethnic group. If the teacher did not understand English well there could be problems.*

The two teachers who did not believe that the assessment was fair across groups of teachers cited possible differences in performance stemming from differences in the quality of preparation related to the type of education received and/or the interview format:

> *One's ability to perform well on the assessment will have a lot to do with the education you received while becoming a teacher. There could be an economic division between those who go to different schools. Also, this assessment could be unfair to teachers who have a shy disposition. They may do well in a classroom where they know the students, but perform poorly with strangers while being taped.*
>
> *Some people may have difficulty using an interview format. Some may be extremely capable, but may not be able to articulate it well in the interview.*

When asked the same question about fairness across groups of teachers, two of the three assessors responding gave affirmative answers, while the third assessor stated, "The assessment does have an undercurrent based for a middle class white situation. Remember undercurrent not tidal wave."

Appropriateness as a Method of Assessment

Teachers were asked if they thought that a semi-structured interview is an appropriate method of assessing competency in teaching secondary social studies. Seventy-five percent of the teachers (12 of 16) agreed, but many qualified their answers. For example:

> *I think if it is refined, and if more day to day routines are discussed, this seems like a useful start. I do feel that many SS teachers (especially me) who are not teaching a particular class may not know the history well enough and it might make more sense to talk about what we are teaching at the time. of course, subject-matter competence is another matter...*
>
> *Insofar that it's an improvement over existing methods. Seems a lot cumbersome, though.*
>
> *Definitely better than multiple choice, but perhaps an additional component of group discussion to assess interaction is appropriate.*

One teacher who answered affirmatively and another who responded, "Yes and no," suggested that this approach seems useful, but only as part of a battery of assessments:

> *Yes, if used in conjunction with other forms of assessment. This showed my ability to deal with content and some strategies. It however does nothing to assess my ability to control a classroom environment or motivate students. No control = no learning. The best laid lesson plans, ideas, strategies, etc. aren't worth anything if you can't provide a good atmosphere. Classroom management, or lack of it I should say, leaves a teacher with content with no one to teach.*
>
> *I do not feel that I did well on most of the tasks, yet I feel I am a pretty good teacher. The training program I completed to be a teacher in no way prepared me for today's assessment. it was very difficult and at times frustrating. I do, however, think that you can learn a lot about a teacher's competency while using this assessment. Most of the tasks were very relevant. It would be most useful when combined with other measurement strategies. Also, you can be a great teacher, but do a terrible job at interviews, so a variety of methods is best.*

The teachers who did not feel that the assessment was an appropriate measure of teaching competency argued along the same lines as the teachers who advocated a battery of assessments, generally stating that a semi-structured interview lacks the ability to tap some aspects of teaching:

> *It might be an indication, but there are many intangible traits not measured. May be used to assess certain skills, but not "competency."*
>
> *It needs a classroom observation component -- especially when the class is not a mainstream class.*
>
> *While each individual college major may differ, our ability to teach that subject cannot be measured by a multiple choice or assessment test. A political science major may be the best history teacher there is, yet he can't pass a multiple choice exam. Given time, he will learn all the simple aspects of the subject. If California wants the best teachers...go after people, not test scores.*

In addition to being asked to comment on the appropriateness of the assessment method, assessors were asked to identify strengths and weaknesses of the assessment instrument. The strengths which they cited were as follows:

- the scope of the tasks which allows the teachers to show their knowledge in a variety of areas and in a variety of ways,
- freedom from local "personality" issues,
- the emphasis on demonstrating skills,
- the allowance for multiple correct answers, and
- an opportunity for teachers to demonstrate factors such as philosophy, caring, and sensitivity to students and content.

One assessor praised the instrument, saying, "[It] honors the teacher by allowing him/her to tell not only what they know and do, but why they do it."

In identifying the instrument's weaknesses, three of the four assessors cited the length of time required to complete the interview. Two assessors noted that not all tasks were applicable to all teachers. One assessor suggested that the instrument "perhaps does not consider the multitude of variables involved in classroom teaching."

Comparison with other assessments. When asked to compare the format of the semi-structured interview assessment with other formats of teacher assessment, roughly 44% (7 of 16) believed it to be better than most other formats, 38% (6 of 16) named another form of assessment which they believed to be better, and 19% (3 of 16) suggested that it provided important information which complemented other assessments.

For the most part, those who believed that the semi-structured interview was an improvement over other formats generally cited its authenticity with respect to teaching skills and/or its comprehensiveness:

> *This assessment format simulates (closely) actual skills required in the classroom.*
>
> *This assessment is more direct and addresses a broader spectrum of teaching abilities than the isolated observation and the dreaded multiple choice tests.*

> *In comparison to multiple-choice tests, this assessment format allows the evaluator to assess a teacher's creativity, thinking, and actual teaching skills.*

Of the six teachers who preferred other methods of assessment, four specifically mentioned classroom observations. One of the two remaining teachers cited the limited topic of the assessment as a severe drawback and the other teacher perceived other approaches to teacher evaluation as providing more extensive feedback. Some of the comments of the teachers who preferred other forms of assessment are as follows:

> *Interviews is a unique format and I believe would give some insight on what type of teacher you are dealing with. But in no way could it completely replace watching somebody in action in the classroom.*
>
> *Classroom observations are very important for evaluation, presence, clarity, rapport, organization, etc. I don't feel these can be accurately measured with the assessment.*
>
> *I would prefer to be evaluated by classroom observations coupled with out-of-class interviews about the same issues (i.e., my rationales for doing what I do). This format depends greatly on teachers being quite articulate on our feet, and having tremendous endurance.*
>
> *Compared to classroom observation, I feel this type of assessment allows a teacher to explain in detail their knowledge of the subject. However, it does nothing to simulate or measure class management.*

Thus, in general, teachers seemed to prefer the semi-structured interview format to multiple-choice examinations, but either preferred observations over semi-structured interviews or believed that some combination of assessment approaches was necessary to capture the full range of essential teaching skills.

Assessment Format

One purpose of the pilot testing of the SSI-SSS was to see if the assessment development methodology for the semi-structured interview in mathematics successfully transferred to a different subject matter area. Therefore, the assessment format becomes of particular interest. In this section, we examine format features, the clarity of preparatory materials, the length of tasks, and examine the evidence about the success of duplicating the assessment methodology in another subject area.

Format Features

This assessment instrument consists of a set of tasks. Each task provides teachers with a set of materials with which they work to accomplish some objective. The teachers are then interviewed with respect to the decisions they made on how to use the materials, and are asked how their decisions might differ with respect to students of other backgrounds, interests, and ability levels.

Clarity of Preparatory Materials

Prior to the assessment, teachers received a general description of the activities involved in each task, together with a sample question that might be asked. They were also asked to bring to the assessment a sample of two student essays, one of which represented an average response, and one of which represented a "less than adequate" response. The essays could be on any topic, but were to reflect a minimum of thirty minutes of writing, either in-class or outside class. The scoring criteria were provided on a card during the initial oral orientation to the assessment.

Teachers were generally satisfied with the information provided prior to the assessment. Eighty-eight percent (14 of 16) of the teachers believed that the assessment activities were clearly described and 75% (12 of 15) believed that the aspects of teaching being measured were clearly described. However, only 31% (5 of 15) believed that the scoring criteria were clear.

When asked if there was any other information that would have been helpful to know prior to the assessment, the following suggestions were made:

- more specificity regarding the topic;
- more direction as to the type of preparation that would be useful; and
- an explanation of the exact teaching skills being observed or monitored.

Clarity of Task Instructions

Teachers were asked if the directions for the tasks were clear, and to identify any problems that they believed resulted from unclear directions. As can be seen in Table 9.4, tasks varied in the clarity of instructions. Although the majority of all teachers found the directions for all tasks to be clear, a significant minority of the teachers experienced problems with **Historical Interpretation** task, and a few teachers experienced problems with others as well.

For the **Historical Interpretation** task, the teachers were presented with three generally contradictory interpretations of the causes of the U.S. Civil War. The instructions included the following direction:

> *Please read the specific selections and then consider approaches you might use to help students understand the different interpretations presented here as well as the concept of historical interpretation in general. For example, you may want to consider the use of activities, analogies, comparisons, etc. Please remember to keep in mind the three factors listed on the evidence card [i.e., the social studies involved; the strategies you use in teaching social studies; the background, needs, interests, and abilities of students].*

Teachers described their difficulties with the **Historical Interpretation** task as follows:

> *I was unclear as to exactly what was expected of me. Coming up with "ways to use material" in the classroom was too vague for me.*
>
> *I did not know what to do for 20 minutes with the Historical Interpretations.*
>
> *Historical Interpretation proved the most difficult to explain in terms of how I would teach H.I.*

Another teacher noted that although the instructions were generally clear, the questions were not because they "were often long and contained several parts" and were delivered orally. This teacher suggested providing each teacher with a written copy of the questions to be read as the interviewer asked the questions orally. FWL staff also observed

TABLE 9.4

TEACHER PERCEPTION OF THE CLARITY OF INSTRUCTION FOR THE SEMI-STRUCTURED INTERVIEW IN SECONDARY SOCIAL STUDIES

Task	# Completing Task and Returning Form	Number and Percent of Teachers Indicating Directions for the Task Were Clear	
		Number	Percent
Unit Planning	16	16	100
Use of Documentary Material	16	16	100
Historical Interpretation	16	11	69
Alternative Pedagogical Approaches	10	10	100
Evaluating Student Learning	15	12	80

teachers having difficulty with questions that had multiple subparts, and recommend that each individual subpart should be asked as a separate question.

In order to communicate the areas in which the teachers would be scored, teachers were given an "evidence card" on which were listed various areas which were to be addressed in their response, e.g., the history and social science concepts, methods of social studies instruction, and student characteristics like gender, ethnicity, and social background. Interviewers occasionally asked the teacher to refer to the evidence card when formulating a response to a question. This tactic had mixed results. Some teachers were visibly confused by the card. Other teachers systematically addressed each area listed whenever they were referred to the card, but if they were not specifically directed to the card, they tended to focus narrowly on the specific question asked, covering only a few of the areas listed on the card.

Length of Tasks

The interviews collectively exceeded the allotted time in a large number of cases, as seven of the seventeen teachers only completed four of the five tasks due to insufficient time. Despite this, some teachers indicated that they would prefer additional time for at least one task. Teachers were asked if they had enough time to complete each of the tasks. Sixty-three percent of those responding (10 of 16) reported the allotted time to be sufficient. Of those identifying tasks for which they had insufficient time, four named **Use of Documentary Materials**, and two named **Alternative Pedagogical Approaches**.

Success in Duplicating the Methodology in another Subject Area

Because the SSI-SSS was not scored, conclusions about the transferability of the assessment methodology from mathematics to social studies can only be tentative. The tasks developed for the mathematics interviews readily transferred for the most part to parallel tasks in the social studies interview. The one mathematics task that needed substantial adaptation was the **Alternative Mathematical Approaches** task. In the mathematics interview, this task consisted of a discussion of the advantages and disadvantages of five different approaches to solving a mathematical problem. In the social studies interview, this task focussed on using different explanations of a single historical event to teach the concept of historical interpretation. The **Historical Interpretation** task consisted of a series of questions exploring teaching a central skill in social studies, historical interpretation. Although most of the tasks in the two interviews were similar, the time it took to complete the interviews was not; there was great difficulty in completing all five social studies tasks in the allotted period of time. One reason that the social studies interviews may have been longer is that the social studies teachers seemed to talk more than the mathematics teachers. If this is the case, reducing the number of questions in the social studies interview may be necessary.

As described earlier in the Scoring section, the scoring indicators in the domains of Content Pedagogy and Knowledge of Students seemed to work well to categorize the evidence collected in the social studies interview. It is likely that the two indicators used for mathematics in the domain of Curriculum Content would be collapsed into a single indicator for social studies because of the centrality to the disciplines comprising social studies of making connections between concepts.

Knowledge in the Curriculum Content domain appears to be more difficult to measure in social studies than in mathematics. In mathematics, it is possible to choose topics containing concepts with which it can be assumed that every teacher is familiar, even if they have not taught them. Certain basic concepts (e.g., the multiplicative identity) will be used in nearly every mathematics class, and ignorance of these principles demonstrates a serious lack of subject matter knowledge. In contrast, there seems to be a lack of "basic" topics in social studies. Social studies consists of a set of disciplines (history, political science, economics), each of which has its own elementary concepts and principles. The social studies entail the understanding of social systems. While the principles may apply across places and times (e.g., countries, regions, historical periods), it is necessary to know a number of specific facts peculiar to the country, region, or historical period to illustrate one's knowledge about the principles, especially to support one's contentions with the depth and specificity necessary to earn higher ratings. However, these details are probably less critical for long term recall in social studies than the principles which they can illustrate. In the context of this instrument, how should a teacher be rated, for example, who can explain the general significance of the Kansas-Nebraska Act but can neither appropriately sequence it in relation to other events of the same era nor supply many details about the specific event? To complicate the evaluation further, some teachers responding to the task will have taught the pre-Civil War era recently while others must struggle to remember their college coursework. Given that the SSI-SSS was not scored, it is not clear how the assessment developers intend to ensure fairness in the measurement of knowledge in the Curriculum Content domain.

The Knowledge of Students' domain appears to be more complex in social studies than in mathematics, possibly because it seems to affect more dimensions of content pedagogy. Discussions of mathematical topics do not typically carry the same affective connotations that discussions of many topics in social studies, especially sensitive topics such as slavery, carry. Furthermore, student backgrounds and experiences seem to play more key roles in student interpretations of historical events than in their interpretation of mathematical problems, requiring social studies teachers to consider student backgrounds when they design instruction. For instance, one teacher observed that slavery was a difficult topic for her to teach because she had students from other countries who grew up under conditions tantamount to slavery and thought being fed and clothed in exchange for freedom was not necessarily a bad bargain, while her Black students bristled at any suggestion that slavery was anything other than reprehensible. Planning instructional activities that can incorporate the probable expression of such disparate viewpoints is a delicate balancing act with no analog in mathematics. Furthermore, reading and writing ability, often thought of as the province of English teachers, as well as reasoning ability, play an important role in the selection of instructional approaches and activities in a social studies classroom, especially when the classroom contains students for whom English is not a native language.

Generally, the assessment methodology transferred well to the social studies domain. Some differences between mathematics and social studies as content areas (e.g., the lack of "basic topics" in social studies) suggest problems remaining to be solved for the social studies interview; other problems identified, e.g., measurement of knowledge of students, suggest the need for continued work on the methodology itself.

Cost Analysis

Complete cost estimates were not possible because of the lack of data on scoring costs. Costs are likely to be somewhat higher than the $137 estimated for administration and scoring of a half-day interview in mathematics.

Technical Quality

Because scoring was not completed, no data were available to evaluate the internal consistency of tasks or to comment on differential group performances as a possible indication of problems in validity.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Semi-Structured Interview in Secondary Social Studies. The section presents information in the areas of administration, scoring, content, format, and concludes with a brief summary.

Administration of Assessment

The Semi-Structured Interview in Secondary Social Studies, developed by the Connecticut State Department of Education, consists of five tasks focused on the teaching of the pre-Civil War era in U.S. History. Teachers are asked to perform a task, then respond to questions about their rationale for the decisions made. The interview is scored holistically with reference to an ordered series of examples of teacher performances corresponding to ratings. The rating is done at the indicator level; with one indicator in the domain of Curriculum Content, and two each in the domains of Content Pedagogy and Knowledge of students. The SSI-SM is modeled after an interview in secondary mathematics previously developed by the same assessment developer.

The semi-structured interview format is complex to administer, requiring careful coordination of tasks to accommodate variation among teachers in terms of the length of interviews. The design of this pilot test minimized transition problems by eliminating intermediate transitions; each teacher was interviewed by a team of two interviewers who divided the tasks between them, eliminating the need to switch rooms. Other administrative issues which were identified in this pilot test include:

- Compared with teachers in other subject areas, beginning social studies teachers appear to be located in lower population densities, suggesting that centralized administration sites for social studies assessments might achieve lower economies of scale and/or require teachers to travel further than centralized assessments in other subject areas.

- The social studies interviews took considerably longer time to complete than did the math interviews; seven of the seventeen social studies teachers were administered only four of the five tasks to keep with the four hours allotted.

- The social studies assessors found their schedule (i.e., five consecutive ten-hour days with one-hour lunch breaks and a half-day break in the middle) to be very demanding, with one assessor describing it as mentally exhausting.

- Many of the participating teachers reported that the collection of student essays was problematic because they did not regularly require their students to write essays. Some teachers had classes with large numbers of students with limited proficiency in English. When requiring samples of corrected student work, it may be necessary to allow for a variety of assessment methods to accommodate the student diversity in California classrooms.

Scoring

The scoring of the assessment was not completed, due to the assessment developer's lack of funds to complete development of the scoring system. The limited work that was done seemed to suggest that the scoring approach previously used for the mathematics interview could transfer to the social studies interview, except that the two indicators of knowledge of Curriculum Content would be collapsed into one for the social studies interview.

Assessment Content

Based on the observations of FWL staff, as well as information collected from assessors and teachers, the following conclusions are offered about the content of the SSI-SSS:

- Although the content of the SSI-SSS was not developed with reference to the *California History-Social Science Framework*, it addresses nearly all of the strands identified under the three major goals listed in the Framework in at least a limited way. Three of the eleven strands receive extensive coverage.

- Of the eleven relevant teaching standards used by the Commission on Teacher Credentialing to evaluate teacher preparation programs, the SSI-SSS best assesses the standard on diverse and appropriate teaching. It partially assesses four other standards which focus on content pedagogy. It provides limited information with respect to two other standards focussing on communication and affect in the classroom, and provides no information with respect to three other standards with varying foci.

- Teachers almost unanimously agreed that the tasks chosen for this assessment were relevant to their job of teaching secondary social studies.

- Roughly two-thirds of the teachers believed that they had sufficient opportunities to acquire the knowledge and abilities needed to respond in a reasonable manner to the assessment questions.

- Almost half of the teachers reported difficulty with the **Historical Interpretation** task, with most citing their unfamiliarity with historical interpretation.

- Teachers overwhelmingly agreed that the assessment is appropriate across grade levels and is appropriate for teachers of diverse student populations. However, the two teachers who did not agree that it was appropriate for teachers of diverse student populations were the two teachers of ESL classes participating in the assessment.

- The assessment, developed for the Connecticut context, embedded some assumptions about the teaching context that were not true of California schools, e.g., that classes were homogeneously grouped with respect to ability, that students could write a coherent essay. Since the SSI-SSS was not scored, it was not possible to see the implications of these different contextual assumptions on evaluating teacher responses. One clear result, however, was the failure of the second half of the **Evaluating Student Learning** task which asked teachers to bring in two samples of essays written by their students. Although nearly all teachers brought examples of student writing on an assigned topic, few teachers brought essays which were well-developed. Many teachers commented that essays were not the most appropriate way to evaluate their students, due to their limited writing skills. We recommend that alternative methods of evaluating student learning be accommodated in any similar task developed in the future.

- The term "average ability" clearly had different referents for the California teachers and the Connecticut assessors, and perhaps had different referents among the California teachers who came from districts with different average levels of achievement. If they are used in an assessment, relational terms such as "average" and "highly capable" need to be clearly defined in ways that communicate a precise meaning.

- The previous two points suggest caution in adopting assessments that were developed in other contexts, and the need for pilot testing with California teachers to evaluate the extent to which an instrument is appropriate in the California context.

- Despite directions to focus on a specific type of classroom (e.g., a classroom of students of average ability), the beginning teachers nearly always responded to the questions with repeated references to their own students. Since this appeared to be a common response across the pilot tests of other assessments as well, it seems that either the assessment must either focus on how a teacher would teach their own students, perhaps elaborating their response to include how they would or not amend their plans to teach other types of students, or include repeated reminders as to the type of classroom that is the focus of the questions.

- Teachers overwhelmingly agreed that the assessment is fair across different groups of teachers.

- Three-quarters of the teachers believed that the semi-structured interview format is an appropriate way of assessing competency in teaching social studies, although only a little less than half believed that it was superior to other assessment approaches.

Assessment Format

The strength of the semi-structured interview format is that it provides for collection of in-depth evidence of specific teaching skills through allowing a teacher to explain the rationale behind teaching decisions. The format especially lends itself to the display of skills in planning and design of instruction, but in terms of assessing the application of skills, it is limited by the lack of evidence of what the teacher actually does in the classroom. Significant findings and recommendations about the format based on this pilot test include:

- Although a majority of all teachers found the directions for each task to be clear, a significant minority of the teachers experienced difficulty with the **Historical Interpretation** task.

- Teachers appeared to have difficulty with questions which contained multiple parts. To avoid confusion, each subpart should be asked as a separate question.

- To reduce the probability that little or no evidence was collected for a particular indicator, teachers were given a card listing areas to be addressed in the responses to questions. This tactic did not consistently work for all teachers, even for questions in which teachers were specifically directed to the card to frame their response. Teacher reactions ranged from confusion on how to use the card to sequential address of each area listed on the card. We recommend that the orientation materials include a section on areas to be addressed in responses together with a sample response which addresses multiple areas.

- Over one-third of the participating teachers reported needing additional time to complete the tasks. Seven of the seventeen teachers failed to complete all five tasks because the question-and-answer period took more time than expected. We recommend that the number of questions be reduced to stay within the time limits.

- Although significant differences between the mathematics interview and the social studies interview were found, it appears that for the most part, the design of tasks and holistic scoring used in the math interview transfer well to the social studies interview. Some problems appeared, e.g., cuing the teacher to display their knowledge of students, which need to be worked out. However, it is possible that these problems are also characteristic of the math interview, but are not as readily apparent.

Summary

The SSI-SSS is still in the process of development, as the scoring system has not yet been tested. Substantial work remains to be done before it could be considered for use in the California context. This work includes shortening the interview, completing the piloting of the scoring indicators, revising some directions and questions to conform to the California context, and refining a strategy to use or to overcome typical beginning teachers' tendencies to focus on their own students. However, despite a few subject-specific problems identified, the semi-structured interview methodology used in the math interview seems to be transferrable for the most part to social studies.

CHAPTER 10:

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

The Assessment of Competence in Monitoring Student Achievement in the Classroom, designed by the Northwest Regional Educational Laboratory (NWREL), consists of a set of ten exercises to which the teachers respond in writing. The assessment is built around a staff development component that provides teachers with instruction on measuring classroom achievement. This instruction had been previously developed by NWREL as part of a decade-long analysis of the task demands of classroom assessment conducted by NWREL (Stiggins, Conklin and Associates, in press). In the pilot test, pre- and posttests were given to both a set of teachers who participated in the staff development activities and to another set of teachers who did not. Two parallel forms of the instrument were developed. The two forms were distributed evenly among the treatment and control teachers for the pretest. For the posttest, each teacher was given the form which they had not taken in the pretest.

Each form consists of ten exercises, each of which use a brief paragraph to describe a specific situation related to the day-to-day monitoring of student achievement in the classroom. Some exercises call for the construction of a particular form of assessment, such as a few items in a paper-and-pencil quiz or a structured observation plan. Others ask the teachers to describe a course of action they would recommend to solve the assessment problem presented. Still others ask for the expression and defense of an opinion about a day to day classroom assessment issue.

The assessment developer identified the following six dimensions of competence in the monitoring of student achievement as the focus of the assessment:

- Understanding of and ability to carry out the full range of uses of classroom assessment

- Understanding of achievement targets for students and the ability to translate those into appropriate assessment methods

- Ability to judge and maximize the quality of soundness of assessments

- Understanding of and ability to use the full range of tools available for classroom assessment

- Understanding of the role of assessment as a dynamic interpersonal activity

- Ability to transform assessment results into sound feedback on performance

The staff development consisted of six three-hour sessions occurring after school, arranged on two consecutive days in each of three months. The topics covered in the staff development, in the order presented, were:

1) Understanding the meaning and importance of high-quality classroom assessment

2) Measuring thinking skills in the classroom

3) Constructing paper and pencil assessments for classroom use

4) Using observation and judgement in classroom assessment

5) Understanding standardized tests

6) Developing sound grading practices

Each topic is addressed by one or more exercises in the assessment.

The assessment is scored through a comparison of the teacher responses to a predetermined set of correct answers. For some exercises, partial points are available for responses which exhibit some, but not all, characteristics of a response deemed to be complete. For other exercises, teachers need to provide only some of the possible responses, e.g., one positive feature of an assessment approach when four specific features are listed as correct.

Administration of Assessments

This section on administration of the assessment contains an overview of the assessment administration, a description of the required logistics, a discussion of security needs, a description of the assessors and their training, and a brief description of teacher impressions of the administration.

Overview

The administration of the assessment occurred in Northern California. Participating teachers came from elementary schools in two districts. The pretests were given after school on either March 22 or March 26, 1990. Staff development activities were then conducted. Most posttests were given after school on either May 22 or May 23, 1990. Seven teachers could not come on the scheduled date for the posttest, so the assessment was administered to them on other dates by district administrators.

As can be seen in Table 10.1, a total of 50 teachers participated in the pilot test, with 33 in the group participating in staff development and 17 in the group which did not. Forty-six teachers completed the posttest. The teachers were distributed almost evenly across the two forms of the assessment.

Table 10.2 shows the characteristics of the teachers in the sample. The majority of the teachers in both the staff development group and the non-staff development group were women. There were four minority teachers in the group receiving staff development and two in the other group. Of the 39 teachers indicating which grade they taught, most taught in the intermediate grades (grades 3-6), although the teachers not participating in staff development were almost evenly split between the primary (K-3) and intermediate grades.

Teachers were given as much time as they needed to complete the assessment. Times for completion ranged from less than an hour to two-and-a-half hours. The median was between an hour-and-a-half and an hour and forty-five minutes.

Logistics

Logistical arrangements included identifying a sample of teachers, administering the assessment (both before and after staff development), arranging for staff development, and acquiring evaluation feedback from participating teachers.

TABLE 10.1

PARTICIPATION IN STAFF DEVELOPMENT ACTIVITIES BY PARTICIPATION IN PRE- OR POST-TESTS

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

	Number of Participating Teachers			
	Pre-test		Post-test	
	Form A	Form B	Form A	Form B
Teachers receiving staff development	17	16	15	16
Teachers not receiving staff development	9	8	8	7
Total number of teachers completing evaluation forms	26	24	23	23
Total # of Teachers	50		46	

TABLE 10.2

PILOT TEST PARTICIPANTS

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Descriptive Characteristics of Participants	Participation in Staff Development Activities	
	Teachers Receiving Staff Development N=33	Teachers Not Rcvg. Staff Development N=17
Gender		
Male	10	4
Female	23	13
Ethnicity		
Asian	1	1
Black	2	0
Hispanic	1	0
Native American	0	1
White	25	8
No Response	4	7
Grade Level		
K-2	10	4
3-6	19	5
No Response	4	8

Identifying teacher samples. The inclusion of the staff development component in the pilot test design made it imperative that teachers be located within a relatively concise geographic area. Districts were contacted about providing groups of elementary teachers to participate. Two nearly neighboring districts agreed to encourage their teachers to participate in the staff development and to solicit a group of comparison teachers who would only participate in the assessment. (The comparison teachers were to receive the training at a later date.) Teachers were paid by their districts for their participation.

Orientation materials. Teachers were given no formal orientation materials; information concerning the assessment and staff development activities was distributed by district administrators, and was chiefly limited to the topic, directions to the assessment and staff development sites, and dates.

Conducting the assessment. The assessment is designed for large-scale administration by a small number of test administrators, who distribute and collect materials and monitor the teachers. No special training or background in assessment is needed, as instructions are designed to be self-explanatory.

The pretests began with a ten to fifteen minute overview of the research design, covering the following topics: (1) the purpose of the pilot test and descriptions of the spring pilot test activities; (2) identification of the assessment developer and distinctions between the roles of the assessment developer and FWL; and (3) the confidentiality and use of the results. Unfortunately, teachers traveling from their individual schools arrived at different times, and several missed the overview at the pretest. At the posttest, teachers were allowed to begin as soon as they desired after arrival. The overview was omitted for the posttest.

Conducting staff development. The six three-hour staff development sessions were given over a two-month time period. They were scheduled for pairs of consecutive days with approximately one month between each group of sessions. The sessions occurred late in the afternoon on school days and were scheduled to allow the teachers enough time to travel to the staff development site after students were dismissed from school. A series of sites approximately half-way between the two districts was located.

Obtaining feedback from the teachers. Evaluation feedback from the teachers was collected through a survey immediately after the posttest.

Security

Basic security precautions such as guarding copies of the assessment instruments and monitoring teachers for collaboration during the test were taken. It is possible that teachers taking different forms of the test discussed the exercises they completed during the period between the pre- and posttests. (The two forms were clearly distinguishable, as each was printed on a different colored paper.) Teachers completed the form which was not used for the pretest as a posttest. Since the teachers could have discussed their forms and questions with others, this is a potential source of biased results. However, there are several reasons to believe that the effect was minimal. First, two months elapsed between the pre- and posttests. Second, teachers were unaware that the same two forms would be used for the posttest. Finally, the assessment did not have any consequences for the teachers, so it is unlikely that they were motivated to make extensive efforts to learn what was on the other form.

Exercises vary in the degree to which they would be susceptible to coaching or memorization of standardized answers so that a teacher could pass the assessment without understanding the underlying principles. Some exercises are more performance-based, e.g., construct three multiple-choice questions based on a given passage, and this type of exercise should be relatively robust to coaching effects since it depends on content that can be easily varied. The responses to some other exercises consist of stating principles for constructing various types of assessments which are relatively content free, and are vulnerable to the memorization of lists with little understanding of the principles or their application.

Assessors and Their Training

Two members of the FWL staff administered the assessment. No training was provided or deemed necessary. If statewide administration of an assessment of this type were contemplated, standardized guidelines for dealing with potential complications, (e.g., a teacher becoming ill during the test) would be needed.

Teacher Impressions of Administration

Over three-quarters (79%) or 33 of the 42 teachers completing an evaluation form believed that the arrangements for the assessment were reasonable. Those who disagreed cited the time of day (immediately after school) and/or distance from their school. Several teachers commented that it was difficult for them to complete a day of teaching, take a two-hour assessment, and then fill out an evaluation form. The following is an example of this type of response:

> *The pretest and posttest are not accurate reflections of my ability. I was just too tired from teaching all day. I would have rather taken them in the morning and had an A.M. sub.*

Other teachers complained that the test was too long:

> *Please note that the pre- and posttests themselves were extremely lengthy, tedious and time-consuming. A shorter version/format would have been appreciated. I frankly grew bored and tired from writing and am not sure my answers reflect the depth of my knowledge in many areas!*
>
> *The testing/pretesting and evaluation forms are unreasonable to expect teachers to fully answer. Too long, too much thought must go into each answer. You have to shorten your test or your participants will become extremely frustrated and disgruntled. I felt overwhelmed, but still answered fully. However, I don't expect others to follow suit.*

Comments about the length of the test were not as prevalent in the pilot tests that took as long or longer to complete, had similar feedback strategies, but were scheduled for Saturdays.

Scoring

The section on scoring describes the process used to score the instrument, the qualifications of scorers, the training of scorers, and their perceptions of that training.

Scoring Process

The scoring system consists of guidelines for each exercise which determine how many points should be awarded for a response. Total points possible vary from 3 to 8 points for each of the ten exercises. Many of the exercises contributing a large number of points to the total score are composed of subparts, which are independently scored. The scoring guide also lists responses to be awarded intermediate points (e.g., one point instead of two for a specific subpart). These responses are deemed to be partially correct but incomplete.

Each scoring criterion identifies responses for each level of credit earned. For instance, in explaining a national stanine score of 4 in language expression, a teacher is awarded the maximum score of two points if the score is interpreted to mean that Helen outscored 20 to 40 percent of the norm group, 1 point if the response says that Helen scored slightly below the average stanine of 5, and 0 points for any other response. While this example has a single response defined as correct for each level of credit, for some exercises, multiple correct responses are identified for each level of credit. The job of the scorer is to match the teacher response with the appropriate level of credit.

Some difficulty in scoring was experienced in that many of the responses described in the scoring guide were written in terms of technical language relating to assessment, using terms like reliability and validity. The beginning teachers responding to the exercises did not tend to discuss any answer in technical terms, so it was often difficult to judge responses, as they looked little like the criterion responses. For instance, one criterion response was "sample performance with a broad array of structured exercises or observations of naturally occurring events". A teacher response that was scored as correct read, "He could videotape the students on various occasions."

One scorer described the problem as follows:

> *I found the language in which the criteria were written to be quite off-putting...The test-designers might consider how to translate technical jargon into common English. It might also help them to see when exercises may not assess what they think they may be assessing.*

Scorers and Their Training

Scorers. Six people completed the scoring training and scored the assessment. All six have experience in assessment development as well as some teaching experience. Two are FWL staff members; the remaining four are doctoral students in the field of education. Three scorers are former math teachers; one is a former science teacher; one is a former English teacher; and one has teaching experience at both elementary and secondary levels, principally in language arts. All four non-FWL scorers completed evaluation forms.

The non-FWL Scorers differed in their opinion as to the level of knowledge about assessment needed to score the assessment. One scorer believed that only a few items required more technical knowledge of assessment than the average scorer not trained in assessment might possess. Another scorer believed that a good working knowledge of assessment and assessment terminology was needed. A third scorer believed that scorers needed to be highly knowledgeable about assessment and current issues in measurement. The fourth scorer had no opinion.

Three of the four non-FWL scorers believed that teaching experience was needed to score the assessment. One scorer stated, "The scorer's knowledge of teaching is more important than her knowledge of assessment. The scorer should be one who is accustomed to 'standing outside of teaching' and reflecting upon it." The fourth scorer did not believe that much knowledge of teaching was needed to score the assessment, although experience in teaching English or reading would be useful for scoring two of the exercises.

Training of scorers. The scoring criteria were designed to require minimal training to score the instrument. The developer of the assessment instrument conducted the training of the scorers, which took about three hours and covered both forms. An experienced teacher completed both versions of the assessment to provide a set of sample responses distinct from those of the beginning teachers to be scored. The training consisted of the trainer reading the prompt, the sample response, and the scoring criteria. He then explained how the scoring criteria should be applied to that response, and asked for any questions from the scorers, which often sparked discussions of how to apply the scoring criteria. This process was repeated sequentially for each exercise.

There was no provision for independent practice in scoring and monitoring of performance before the scorers began evaluating actual teacher responses. Some of this occurred informally, in the form of individual conversations and informal group discussions about how to apply various scoring criteria.

Only one example of a teacher response for each exercise was used to demonstrate the application of the scoring criteria. Furthermore, the example was from an experienced teacher who had attended a graduate-level course taught by the assessment developer. Consequently, the sample response was much more lengthy and frequently used the technical language of the scoring criteria, unlike the responses of the beginning teachers, which tended to be brief, and couched in general terms. The training would have been strengthened considerably by the demonstration of both a greater number of examples of scored responses and by providing examples which more closely resemble those to be scored.

Perceptions of training. All four scorers who were not FWL staff completed evaluation forms, which included questions on the training. All four of the scorers rated the training as "adequate," the intermediate rating provided. However, one scorer added the qualifier "barely," and another noted that "I did not feel sufficiently confident that I was applying the scoring criteria like other scorers." The scorers agreed that the most useful part of the training was the discussion of specific examples of applications of the scoring criteria. Two scorers recommend adding this to the training. The other two recommended more extensive preparation for scoring, either by taking the assessment prior to scoring it to learn its contents or by reviewing materials sent in advance of the training which explain "what the test is designed to do, who it tests, under what conditions was the testing done, etc."

Assessment Content

This assessment differs from the others pilot tested for the California New Teacher Project in that it does not focus on a single subject matter, but on a teaching competency which cuts across subjects: monitoring the achievement of students in the classroom.

The importance of the topic of classroom assessment is supported by an informal survey of the teachers attending the initial staff development session. When asked if they had received any training in assessment during their teacher preparation, only a few responded affirmatively. A growing number of regional and national surveys indicates that it is typical for teachers to lack training in assessment. Yet all teachers are expected to employ both formal and informal assessment techniques and to make judgements about students.

The assessment developer has spent the last decade in the designing, testing, and redesigning of instruction in assessment for classroom teachers, and is a nationally-recognized authority in this area. The topics selected for the staff development component of the assessment are a subset of the instructional modules he has developed. This assessment is intended to test knowledge and application of principles for sound construction and proper use of student assessments that are based on both research and teacher feedback.

In the following pages, the content of the Assessment of Competence in Monitoring Student Achievement in the Classroom is evaluated along these dimensions:

- Congruence with various curriculum frameworks addressing curriculum in the elementary grades;
- Extent of coverage of California Standards for Beginning Teachers;
- Job-relatedness of the instrument;
- Appropriateness for beginning teachers;
- Appropriateness across different teaching contexts (e.g., grade levels, subject areas);
- Fairness across groups of teachers (e.g., ethnic groups, gender); and
- Appropriateness as a method of assessment.

As was true of all of the assessment instruments pilot tested this spring and summer, there was not sufficient time during development to conduct a larger content validity study. Without such a study, our ability to comment on the assessment's appropriateness along such dimensions as job-relatedness, appropriateness for beginning teachers, and appropriateness across contexts is limited. Thus, excluding the first two dimensions of curriculum congruence and standards coverage (which are based on FWL staff's analysis of the documents involved), the discussions of the remaining dimensions are based on the perspective of the participating teachers and scorers, and of FWL staff, as reflected in feedback forms, in informal conversations with the scorers, and in analysis of the scores.

Congruence with California Model Curriculum Guides and Frameworks

The following discussion of the content begins with a comparison of the assessment instruments with the model curriculum guides. This assessment emphasizes knowledge of principles of valid assessment which pertain to every curricular area. Assessment is not typically addressed at length in the curriculum guides and frameworks, which mainly focus on curriculum content. However, there are aspects of nearly every curriculum guide and framework which address evaluation of student progress. The particular instruments pilot tested included references to evaluation of student progress in four subject matters: language arts, science, social science, and mathematics.

Language arts. Three exercises portray assessment in the subject of elementary language arts. One exercise on each form addresses writing assessment, and an additional exercise on one form addresses the assessment of reading. One of the writing assessment exercises asks a teacher to give a student written feedback on a writing sample for several features chosen by the candidate (excluding mechanics). The focus of this exercise is on the teacher's ability to devise appropriate criteria, apply them, and explain their evaluation to a student in a way that provides useful feedback. The writing sample used for this exercise is an account of a student's friendship over time. This is consistent with the emphasis in the *English-Language Arts Model Curriculum Guide: Kindergarten through Grade Eight* (California State Department of Education, 1988) on basing instruction on students' experiences.

The second writing assessment exercise, on one form only, asks the teacher to list some of the features of a writing assessment designed to provide diagnostic information and show change over time. However, some of the acceptable responses for this exercise are consistent neither with current research on writing nor with the emphasis on basing instruction on students' experiences in the *English-Language Arts Model Curriculum Guide.* In this exercise, two scoring criteria listing features of a writing assessment ("sampling with sound writing prompts" and "keeping the prompts constant over time") both suggest that the topic for writing comes from the teacher. One of the key recommendations in writing instruction is the importance of writer-generated topics (Graves, 1983). Another scoring criterion for the same exercise, which requires concealment of the writer's identity to avoid bias, also runs counter to writing research which stresses the influence that background knowledge plays in a reader's or writer's construction of meaning. Familiarity with this background knowledge is deemed necessary for effective evaluation of student work. The preselected topic and concealed identity of the writer might make more sense for large scale program assessment, but it is less appropriate for classroom assessment.

The third exercise, on one form only, portrays three approaches to evaluating reading ability. Teachers are asked to identify positive and negative features of all three approaches. One approach measures good reading by the ability to answer paper-and-pencil questions about what was read, i.e., the ability to construct meaning from a text. Another approach tests for the ability to read fluently, i.e., sound out words, and the third examines the ability of readers to retell the meaning of what was read in their own words with fidelity to the true meaning, i.e., reproduce the message in the text.

To the extent that the first approach is linked with the completion of narrowly-focused worksheets, it conflicts with the emphasis in the relevant Model Curriculum Guide cited previously. The Model Curriculum Guide emphasizes that students should be encouraged to actively interpret texts and that such interpretations vary from student to student, as each student constructs meaning based on their own set of background experiences. The second approach would not be considered by most reading experts to be a good single measure of reading ability, and is not mentioned in the *English-Language Arts Model Curriculum Guide*. Research on reading shows that while the ability to sound out words is often associated with the ability to construct meaning from the text, sometimes readers can sound out words without understanding the meaning. The third approach is consistent with the emphasis in the *English-Language Arts Model Curriculum Guide* on an interrelated program of listening, speaking, reading, and writing. To the extent that it implies that meaning inheres in the text independently of the reader, however, it conflicts with the *English-Language Arts Curriculum Guide*.

Science. Four exercises across the two forms use science as the focal subject of assessment. These four exercises constitute two sets of parallel exercises on each form. One set asks teachers to list principles for evaluation of the potential of either multiple-choice unit tests provided with textbooks or laboratory activities as assessment instruments. The other set asks teachers to construct multiple-choice items testing both recall and higher order thinking skills based on a given passage from a science textbook.

The latest science framework available is the *Science Framework for California Public Schools, Kindergarten Through Grade Twelve* (California State Department of Education, 1990), which was released after this assessment was developed. As the most up-to-date statement of desirable content and framework of the California science curriculum, however, it is the standard to which the assessment exercises using science content are to be compared. *The Science Framework* supports the goal of increasing time devoted to hands-on activities in science classes to at least 40 percent of the total time devoted to teaching science. The exploration of laboratory activities as tools of assessment could be useful in measuring student achievement during hands-on activities.

Another emphasis in *The Science Framework* is the teaching of science in depth rather than superficially. The evaluation of multiple-choice tests to assess content knowledge gained through study of a unit, is a valid assessment approach to consider. However, the exercise would be more congruent with The *Science Framework* if it were amended slightly to make clear that the textbook was accompanied by a series of laboratory exercises, and the unit tests measure learning from both sources.

The design of multiple-choice tests to assess higher-order thinking skills was one of the staff development activities. However, the excerpts from science textbooks on which the students were to be tested were too brief and superficial to support higher-order thinking without making many assumptions about the background knowledge of students. Furthermore, the testing of knowledge based on a brief passage conflicts with the emphasis in *The Science Framework* on in-depth knowledge gained at least partially through observation or experimentation which goes well beyond information presented in a textbook. An alternative which might be more congruent with *The Science Framework* would be to provide brief descriptions of a small series of laboratory activities together with the scientific principles to be inferred or reinforced, and to ask the teachers to construct multiple-choice questions to test mastery of these principles.

Social science. There were two exercises which focussed on social science content, one on each form. One exercise asked teachers to compare two different assessment approaches: an oral response comparing two countries on a given dimension and an item on a written test comparing the two countries. As with other curriculum frameworks, *The History-Social Science Framework* (California State Department of Education, 1988) emphasizes in-depth understanding of topics as opposed to more superficial knowledge. It also emphasizes understanding the significance of characteristics of governments or countries. The task portrayed in this exercise closely resembles memorization of isolated facts, and could be revised to portray an activity more congruent with the current framework, such as comparing the factors contributing to the evolution of Mesopotamia, Egypt and China as societies (California State Department of Education, 1988: 61).

The other exercise with social studies content asked teachers to suggest alternative ways of assessing limited-English-proficient students who might not understand the multiple-choice, true/false, and fill-in-the-blank items in the unit tests provided with the textbook. *The History-Social Science Framework* calls for more than the assessment of student progress in learning knowledge. Additional goals recommended are: (1) the assessment of basic skills and abilities, including those of thinking and social participation; (2) the utilization of a variety of evaluative techniques, including the teacher's evaluation of the students' performance, students' evaluation of personal progress, and peer evaluation; and (3) opportunities for students to make oral and written reports in which they are encouraged to state a position and support it. While carefully constructed paper-and-pencil tests such as those referred to in the exercise can measure some higher order thinking skills with respect to content knowledge, it is more likely that other forms of assessment would be needed that would likely be more appropriate for both limited-English-proficient students and other students as well, given the additional recommended evaluation goals in the area of social science. For instance, teachers might be asked to design a performance-based assessment to meet one of the goals described above for a heterogeneous classroom that included several limited-English-proficient students.

Mathematics. Only one exercise on one form focused directly on assessment in mathematics, and that exercise portrayed a classroom interchange between the teacher and several students in which the content plays a relatively minor role. *The Mathematics Framework* (1985) emphasizes problem solving and an increase in the use of cooperative learning groups. Some aspects of assessment relating to these emphases were addressed by the exercises, though using other subjects, such as conducting observations and performance assessments and constructing multiple-choice questions to measure higher-order thinking skills. However, problem solving and the use of cooperative learning groups, though not peculiar to mathematics, each pose problems for assessment of student achievement which are not explored by this assessment. How would one measure problem-solving ability, especially when the type of problem solving advocated in *The Mathematics Framework* includes encouraging students to follow incorrect strategies to learn for themselves how to determine when a strategy is not working? How do you individually evaluate students when they are engaged in a group activity? Neither of these assessment dilemmas were addressed by the assessment.

Examination of specific subject-matter content suggests that modifications of the exercises are necessary to bring the assessment into closer congruence with current curriculum guides and frameworks with respect to the content represented in the prompt materials. With respect to the variety of assessment approaches represented in the exercises, the assessment addresses knowledge of how to construct effective performance assessments, observation protocols, and paper-and-pencil assessments. As can be seen from this list, most if not all of the assessment methods which might be used to assess the more in-depth knowledge called for in the latest curriculum guides and frameworks are represented in the current collection of exercises. Sometimes the exercises stop short of measuring a teacher's performance in constructing and using these assessment approaches, testing instead knowledge of general principles of assessment design which the candidate may or may not be able to apply correctly. Eliciting more performance-based responses from the teachers or asking teachers what conclusions they would draw from different assessments that vary in soundness of design would require more direct demonstration of assessment skills.

Extent of Coverage of California Standards for Beginning Teachers

The California Beginning Teacher Standards are criteria for teacher competence and performance which the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. The usual practice in evaluating assessments pilot tested is to consider the stimulus materials and scoring criteria in light of each standard. However, this assessment was narrowly focussed on a single area of teacher competence: measuring student achievement in the classroom. Therefore, it will be discussed with respect to the only applicable standard, Standard 27, the text of which is printed below in italics.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.*

Although evaluating a teacher's ability to use assessment techniques for student diagnosis and evaluation is the goal of the assessment instruments, the exercises are uneven in their ability to accomplish this goal. Some of the exercises (e.g., one which asks teachers to write multiple-choice items to test comprehension of a paragraph of text provided) require teachers to apply their knowledge of general principles related to assessment to specific situations. Other exercises, though utilizing classroom-related problems, only ask the teacher to respond in terms of general principles for constructing a valid assessment, and do not determine whether or not the teacher can apply these principles.

One example of a missed opportunity to test application of assessment-related knowledge is the exercise where teachers interpret standardized test scores. With one exception, the exercise focuses on evaluating the technical accuracy of the explanations of various scores. The exception is where the teachers are asked if they should use a grade-equivalent score to choose the level of work for a child. Teachers could have been asked how they would respond to questions from parents about the meaning of the scores, and judged on whether they could communicate the appropriate meaning.

The major focus of the exercises is on assessment issues, sometimes to the exclusion of other issues in a complex situation with many competing goals. Teachers rarely use such a singular focus to analyze problems. The exercises vary in the extent to which they accommodate competing concerns which might lead a teacher to use less effective assessment techniques to achieve other goals. For example, one set of exercises asks teachers to label assessment practices either "sound" or "unsound" and explain their position.

The criteria for scoring one subpart allows a practice which is otherwise sound from a strict measurement point of view to be labeled unsound due to the effect on the self esteem of students. In contrast, other exercises, e.g., an exercise on how a teacher handles cheating, require teachers to focus solely on measurement effects and disregard any other effects, e.g., penalties for cheating that may affect measurement of achievement but serve as a powerful negative reinforcement discouraging future cheating. (Several scorers also commented that many schools and districts have an official policy on penalties for cheating which teachers are required to follow, and that it would be difficult for beginning teachers to contradict local policy.) If a teacher is directed to evaluate a decision which has potentially negative effects on goals unrelated to assessment, then the instructions should instruct the teacher to consider the effects on assessment alone as an initial step toward identifying the consequences of that particular decision.

Some of the scoring criteria are debatable even in measurement terms, e.g., one exercise which asks teachers to judge whether dropping a student's lowest grade is a sound or unsound method of assessing student achievement. One line of thought, reflected in the current scoring criteria for the highest value, would argue that dropping a grade reduces the scope of the assessments or amount of information over which the student is evaluated, thus reducing the validity and reliability of the assessment. Another line of thought holds that reliability is increased when outliers are eliminated, so the highest and lowest scores could be dropped. And, again, measurement of noninstructional objectives is not addressed in this current version. For example, the motivational objective of deleting or substituting for a low score is not acceptable within the current scoring guide.

Because of the problems described above with the exercises developed, FWL staff judge this assessment to only partially cover Standard 27 of the Beginning Teacher Standards.

Job-Relatedness

Both teachers and scorers were asked whether the scenarios chosen were relevant to a teacher's job of monitoring student achievement.

Teacher perceptions. A total of 42 teachers completed an evaluation form responding to various questions about the assessment. While all teachers completing the form had taken both forms of the assessment as either a pre- or posttest, half of the teachers had taken Form A as the posttest, and half Form B. It is likely that the form used as the posttest was the major influence on their response, as the pretest had occurred two months previously. Therefore, the results are reported separately by form.

When the teachers were asked if they felt "the scenarios chosen for this assessment are relevant to your job of monitoring student achievement," 71% (15 of 21) of the teachers completing form A as the posttest and 67% (14 of 21) completing form B responded affirmatively. Those who agreed did not elaborate on why they believed the assessment to be relevant. Those who disagreed gave several reasons, including type of students taught and grade level. The following comments are illustrative:

> *Most of my students (85% are Chapter 1 students) they need hands on teacher directed assessments. Rarely do I use a standardized test or essay, more of a discussion and an assignment.*
>
> *As a kindergarten teacher it was a bit of a reach to answer the upper grade questions. There were none at my level of teaching experiences.*

Scorer perceptions. All four scorers believed that the assessment was relevant for teachers. "Most of the items present situations commonly confronted by beginning teachers," commented one scorer. However, even while agreeing that the assessment was, on the whole, relevant, one scorer expressed reservations about a few exercises, and another was not certain that the assessment was relevant for kindergarten teachers.

Generally, both teachers and scorers believed that the scenarios represented in the assessment were relevant to the job of elementary teaching.

Appropriateness for Beginning Teachers

The appropriateness for beginning teachers was assessed through both surveys of teachers and scorers and through analysis of the teacher scores.

Teacher perceptions. Teachers were asked if, as new teachers, they felt that they had "sufficient opportunity to acquire the knowledge and abilities needed to respond in a reasonable manner to the assessment questions." Sixty-seven percent (14 of 21) of the teachers completing form A as a posttest and 48% (10 of 21) of the teachers completing form B believed that they had such opportunities. Roughly half of those who disagreed that the test was appropriate for beginning teachers received the staff development training, and half did not. Those who did not attend the training generally cited their lack of instruction in assessment, as in the following comments:

> *I have received no training in assessment. Bloom's taxonomy of types of questions has been the only information I have received concerning assessment.*
>
> *Not one of my education courses covered the rationale behind test procedures (I have a Masters in Instructional Leadership).*

Those who did participate in the staff development workshops gave several reasons for their belief that their preparation had been inadequate, including a need for more classroom experience, and a desire for more practice in application of the principles learned. The range of comments is reflected in the following set:

> *It takes a very long time to practice being reasonably confident in assessment.*
>
> *Several questions require experience as well as a theoretical application. Also even though certain practices were covered in course work such as testing evaluation, they were not covered in depth with practical application.*
>
> *I need time and practice. I liked the workshops but I need more hands-on. I learn by doing.*

The issue of appropriateness for beginning teachers was also explored through another question which asked the teachers if they found any questions on the posttest too difficult. Sixty-seven (14 of 21) of the teachers taking form A and 43% (9 of 21) of the teachers taking form B reported that there were questions they found too difficult. Table 10.3 shows the topic of the exercises the teachers reported finding difficult. The largest number of teachers reported difficulty with interpretation of standardized test scores, particularly with explaining a stanine score. The next largest numbers had difficulty with writing assessment and constructing multiple-choice items.

TABLE 10.3

TOPICS OF EXERCISES REPORTED BEING TOO DIFFICULT

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Topic	Number of Teachers Reporting Topic Difficult
Interpretation of standardized test scores	10
Writing assessment	5
Constructing multiple-choice items	4
Calculating composite test scores	3
Assessing reading	1
Assessment during instruction	1
Constructing an observation assessment	1
Effect of penalty for cheating on soundness of assessment	1

Scorer perceptions. Stating a consensus that most teachers do not presently receive sufficient instruction in monitoring student achievement, none of the scorers believed that typical beginning teachers have the knowledge necessary to perform well on the assessment. However, as one scorer put it, "These are issues that new teachers could be taught to address. They are not necessarily dependent on the amount of experience that one has with students." The scorers see this as a topic that is not beyond the grasp of beginning teachers if they were to receive more instruction in classroom assessment.

Performance on assessment tasks. Table 10.4 summarizes the performance of teachers participating or not participating in staff development. (Since each teacher was scored by two scorers, both scores were summed to arrive at a final score.) Teachers taking Form A as a pretest took form B as a posttest, and vice versa. The highest score possible on each form was 106. As the table indicates, the majority of scores are well under 106, suggesting that the teachers were not sufficiently prepared in classroom assessment to do well on this instrument.

Teachers as a group did particularly well on the exercise which asked them to identify alternative ways to assess LEP students and on the exercise analyzing a teacher's verbal responses to students. They did particularly poorly on constructing multiple-choice questions, suggesting ways to accurately assess students whose cultural norms interfered with common ways of assessing student performance, interpreting standardized test scores, and evaluating specific assessment practices, including a proposed penalty for cheating.

These results do not provide evidence that the staff development had a strong impact on teacher performances on the assessments. The assessment developer has conducted training of similar content in settings where experienced teachers participated in the training in either university courses or district inservice programs. Possible explanations for the absence of clear training effects in the pilot test include (1) given that these were new teachers, inexperienced in not only student assessment but also in conducting their own classrooms, the format of one-session-to-a-topic staff development workshops conducted after school was probably insufficient to completely address such a complex topic in a manner that the beginning teachers could absorb, (2) training where teachers self select an area, which was not the case here, will result in greater impacts than when there is not an explicit commitment by the teacher participants, and (3) the time period, after school setting, and related fact that teachers did not necessarily practice or apply the concepts contained in training sessions would reduce the effectiveness of the training. (In university and district inservice sessions the teachers are given and complete assignments that build on the training provided. The pilot test did not have this feature.)

TABLE 10.4

TEACHER PERFORMANCE BY FORM, PRE- OR POST-TEST, AND PARTICIPATION IN STAFF DEVELOPMENT

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

	Form A		Form B	
	Pretest	Posttest	Pretest	Posttest
Teachers participating in staff development				
Mean	39.82	44.93	45.06	47.13
Standard Deviation	11.95	16.32	11.97	12.23
N	17	15	16	16
Teachers not participating in staff development				
Mean	43.56	41.86	43.50	49.50
Standard Deviation	17.35	14.31	8.75	10.01
N	9	7	8	8

Appropriateness across Contexts

The contexts explored on the evaluation form included both differing grade levels and differing types of students.

Grade level and subject matter. Table 10.5 shows the grade levels represented in the set of exercises for form A and form B. Both forms, but particularly form B, were more weighted toward questions focusing on the highest grades covered by the elementary credential.

Teachers were asked directly if they felt that "the assessment was appropriate for different teachers teaching grade levels from kindergarten to eighth grade." Forty-three percent of the teachers (9 of the 21 teachers taking each form as a posttest) believed that it was. Teachers who did not believe it was appropriate made the following comments:

> *I did not feel comfortable answering exercise 9 [design of an observation assessment to determine whether or not a student should be retained in kindergarten] since I'm not familiar with kindergarten.*
>
> *This assessment does not relate to my daily events and situations -- I was asked to be a junior high teacher, a fifth grade teacher, seventh grade teacher, a sixth grade teacher, but only once asked to be a kindergarten teacher (which I am). Only half of the questions were circumstantially general enough for me to be comfortable answering.*
>
> *The questions were geared toward the middle school grades.*

This last set of comments echoes similar comments on other assessments, and illustrates a paradox that teacher assessments for licensure must address. On the one hand, teachers are licensed to teach multiple grade levels, so situations portraying various grade levels should be represented in the stimulus materials to which teachers are asked to respond. On the other hand, the rationale for delaying administration of performance-based assessments until the early years of teaching is to focus on competencies which take may some unspecified amount of independent teaching experience to develop. The experience of beginning teachers, however, typically only covers a limited number of grades. How to

TABLE 10.5

DISTRIBUTION OF SCENARIOS IN EXERCISES ACROSS GRADE LEVELS

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Grade Level	Number of Exercises	
	Form A	Form B
K–2	2	0
3–5	1	4
6–8	5	4
Unspecified	2	2

BEST COPY AVAILABLE

design assessments so that a teacher can draw upon their teaching experience yet sample the entire range of grade levels covered by a credential is a problem that remains to be solved, particularly by assessments of elementary teachers where not only multiple grade levels but also multiple subjects are included.

Scorers were not asked to comment on fairness across grade levels.

Diverse students. Another dimension of the appropriateness of the assessment across contexts is fairness across differing groups of students taught. California students are increasingly diverse. Only four of the 42 teachers responding to the evaluation survey reported that none of their students spoke any language other than English. Six teachers reported that their students collectively spoke five or more languages. (It should be noted that this does not necessarily mean that students who spoke a language other than English did not speak English fluently, only that California classrooms are increasingly diverse in terms of student cultures.)

The assessment was consciously designed to represent a variety of students, and to place members of gender and racial/ethnic groups in non-traditional roles whenever possible. Examples of the latter are the portrayal of a man as a kindergarten teacher and a black female as the highest scorer on a series of tests. The exercises were examined by the development team to verify that both positive behaviors and negative behaviors portrayed in the exercises were distributed across students of various backgrounds. While most exercises did not discuss the classroom composition, the names used to identify specific students were characteristic of a variety of ethnic groups. One exercise on both forms addressed the assessment of students whose cultural norms might interfere with traditional assessment measures. Another exercise on one form dealt with the assessment of limited-English-proficient students. A third exercise addressed the design of an observation assessment to identify students who should be retained in kindergarten.

Teachers were asked if they believed the assessment was "appropriate for teachers of diverse student groups (e.g., different student ability levels, different ethnic groups, handicapped or limited English students, different school/community settings)." They overwhelmingly believed that it was, with 90% (or 19 of the 21 teachers completing each form) replying that it was appropriate. The one teacher disagreeing who elaborated on the choice cited the exercises on paper/pencil examinations, explaining that Limited-English-proficient students did not take such examinations.

Scorers were asked to comment on the appropriateness of the assessment to address "the ability of the new teacher to work with diverse student groups." Their responses were mixed. One scorer believed that it was appropriate, while the other three expressed varying degrees of reservation. One scorer praised the exercise which asks teachers to suggest alternatives to paper-and-pencil tests to assess limited-English-proficient students, but believed that another exercise focusing directly on how to avoid cultural norms interfering with assessment practices needed to be "overhauled." Another scorer recommended either a greater emphasis on cultural diversity or heavily weighting the same exercise addressing cultural norms which the previous scorer criticized. The remaining scorer was "not sure that either form adequately addressed the issue of diversity," believing that the exercise on cultural norms was "presented so badly that candidates didn't have much of a chance to display their knowledge about how to deal with a multicultural classroom." In informal discussions, the scorers pointed out that many of the teachers discussed the examples in terms of psychological rather than cultural explanations (e.g., the student who did not wish to draw attention to him/herself was perceived as being shy rather than as coming from a culture which stressed communal rather than individual achievement), and several stated their belief that the examples provided misled the teachers.

Scorers were also asked their opinion of the suitability of the assessment for new teachers in different school and community settings. Two scorers believed that it is "general enough" to be used for teachers in different settings, although one qualified the answer by suggesting additional scoring criteria for one exercise. The additional criteria accepted a currently ineligible response if qualified due to an inadequate supply of materials or an extremely large class size. The other two scorers expressed reservations due to differing philosophies of assessment or policies regarding assessment practices which they did not see equally reflected in the scoring criteria.

To further explore the freedom from bias against teachers from particular teaching contexts, the assessment developer's description of the bias review, Form B, and the relevant scoring guide were sent to Dr. Sharon Nelson-Barber, an assistant professor of anthropology at Stanford University, for review. Dr. Nelson-Barber is a consultant who works with school districts and teachers of homogeneous minority classrooms.

Dr. Nelson-Barber praised the portrayal of equity and diversity in the exercises, but she expressed concern about the exercises' ability to do the following:

- accommodate alternative conceptions of teaching

- account for the assumptions about context and preferred teaching practices that guide candidate responses and scorer ratings

- accommodate answers that might be deemed correct by the scorer but do not appear in the scoring guide

The first two concerns do not only apply to this assessment, but to all the assessments which Dr. Nelson-Barber reviewed. Her concern about the ability of the assessment to accommodate different conceptions of teaching centered around the fear that the instructional techniques and interactive behaviors deemed effective in minority communities would not be seen as appropriate. Dr. Nelson-Barber points out that there are many ways to be a good teacher, and she believes that teachers should be evaluated with regard to the ways that they are trying to use.

Both the teachers responding to the assessment exercises and scorers have philosophies of teaching and preferred teaching practices which undoubtedly influence their judgements. Teachers in this assessment are not permitted an opportunity to communicate assumptions about the teaching context in which the incidents portrayed in the exercises occur and how the context affects the reasoning that underlies a particular response. Dr. Nelson-Barber has reservations about scoring a response as correct or incorrect without understanding the contextualized reasoning which led to the response. For example, the exercise designing an evaluation of a writing sample explicitly tells the teacher to exclude mechanics, yet many inner city parents and teachers view mechanics (e.g., grammar, spelling) as extremely important for their students to master in order to successfully compete for good jobs.

FWL staff note that there is little contextual information about the classrooms or districts portrayed in the exercises, so it is likely that teachers draw from their own experience in answering the questions. Teachers of classrooms which require certain assessment strategies, e.g., a large number of ESL students who have difficulty reading English, may find exercises difficult which focus on assessment methods which are not suitable for their students.

Since Dr. Nelson-Barber is not an expert in assessment, her ability to comment on the appropriateness of the range of acceptable responses reflected in the scoring guide was limited. Based on her general experience working in diverse classrooms, she believed that reasonable limits had been set for the scoring criteria; however, she expressed discomfort with the fixed nature of the acceptable responses which does not allow scorer discretion to give credit for acceptable responses that do not appear on the scoring guide.

Fairness across Groups of Teachers

The teachers were asked if they believed that the assessment was fair "to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers. Nearly all of the teachers believed that the assessment was fair, with 90% (19 of 21) of those completing form A as a posttest and 95% (20 of 21) of those completing form B agreeing that the assessment was fair.

The scorers responded to a similar question on their evaluation form. Two scorers commented that the assessment seemed to favor teachers with strong written, as opposed to oral, communication skills. One of these scorers also commented that teachers' subject matter backgrounds could either give them an advantage or serve as a distraction for different exercises. A third scorer believed that teachers with limited proficiency in written English were handicapped, especially in writing multiple-choice items. (However, an implicit emphasis on English proficiency is probably appropriate, unless a teacher is a bilingual teacher primarily instructing in another language.) The fourth scorer qualified the answer as follows: "If the expectation is that all teachers should be able to take and pass pencil and paper tests and that all new teachers have been taught the 'lingo' of psychometrically valid assessment, then it is a fair test." (This scorer had expressed the belief that the exercises did not match what good teachers actually do in their classrooms to monitor student achievement, and that people with the appropriate vocabulary and mastery of abstract concepts could do well on the assessment without being able to actually apply any of the principles they could describe. If the scorer's belief is true, this could be addressed by the inclusion of more performance-based items.)

Appropriateness as a Method of Assessment

Teachers were asked to comment on the appropriateness of the assessment methodology in two ways: by agreeing or disagreeing that they believed the assessment was an appropriate way to assess competency in monitoring student achievement and also by comparing the assessment method with other assessments with which they had been evaluated, with the CBEST, the NTE, and observations during student teaching given as examples.

Teacher perceptions. Thirty-eight percent (8 of 21) of the teachers completing form A as a posttest and 48% (10 of 21) of the teachers completing form B believed that the assessment was an appropriate way of assessing their competency in evaluating student achievement. Those agreeing that it was appropriate gave responses and suggestions such as the following:

> *It makes you think through and explain your positions on assessment.*
>
> *But an "open-book" format would be more useful and accurate -- showing how the competency is used, not memorized.*

Teachers who disagreed questioned the validity of the assessment:

> *Even if I answered everything correctly, it doesn't necessarily mean I have bought into the values expressed and have initiated them in my class. (In my case, however, I am doing my best to improve).*
>
> *The real test is how we assess in our classrooms on real kids that we know and understand. Many of the exercises do not apply to a primary teacher, such as test scores, and multiple choice test items.*
>
> *I think a better way would be to look over my grade book, tests used, and have someone offer practical suggestions as to my method of assessment. (How much should a test weigh compared to daily work? Should curve or straight percent be used? How do you grade a writing sample of a RSP student in a regular reading class compared to others?)*

Scorer perceptions. Scorers were asked to describe the strengths and weaknesses of the assessment. Three of the scorers praised the topic of the assessment. Two scorers saw the ease of administration as a strength. One scorer cited the relative ease of application of scoring criteria. One scorer summarized the strengths of the assessment as follows:

> *I find the general thrust of this assessment instrument valuable. It asks teachers to justify their approaches to assessment in specific contexts commonly confronted in the classroom, and to perform basic tasks related to assessment that all teachers should be able to do. I appreciate the inclusion of questions pertaining to culturally-based beliefs, and the assessment of writing.*

In terms of weaknesses in the assessment instruments, three of the four scorers cited poor instructions and/or stimulus materials, such as instructions which did not always indicate the scope of the response reflected in the scoring criteria, badly written passages from which to construct multiple-choice questions, and ambiguous or poorly worded scenarios. Two scorers criticized the range of responses deemed appropriate, with one commenting that it seemed as if drawing on prior classroom experience negatively impacted the teachers' scores. Two of the scorers also questioned the sample of assessment activities and whether it represents what teachers should know or use in actual classroom practice. Finally, one scorer each mentioned as a weakness the emphasis on principles of assessment instead of their application, arguable scoring criteria, and the particular representation of "good" practice in the exercises focusing on literacy.

Comparison with other assessments. About 38% (8 of 21) of the teachers completing form A as a posttest and 57% (12 of 21) of the teachers completing form B compared this assessment format favorably with other assessment formats with which they had been evaluated. They appreciated the opportunity to explain their answers instead of having to choose among fixed options in multiple-choice questions, as exemplified by two teachers' comments:

> *It better tests one's overall ability to communicate ideas and beliefs than the CBEST or NTE. Multiple choice tests are very limiting and there is always the guess factor to consider. That possibility is ruled out here.*
> *It's really difficult for me to say -- if I liked the questions better I would say that this is a much better assessment than multiple choice tests because I can put my thoughts down for you to see rather than just marking a box answer that someone else put down for me to choose from.*

Some teachers preferred a combination of assessment formats, such as the teacher who suggested "a true blend of CBEST, observation, [and] essay." Two teachers mentioned a preference for classroom observations.

The scorers were asked to identify the unique contribution of the instrument to the assessment of elementary teachers, compared with other assessment methods. One scorer expressed qualified approval of the method employed:

> *This method is clearly preferable to one-shot classroom observations. It does not capture how teachers conduct assessments or use them in their classrooms, as portfolios could. It does attempt to establish general,*

> *standardized questions all teachers should be able to answer in the area of assessment, which is a tremendous advance over little or no concern given to general issues in assessment faced by teachers.*

One scorer had no opinion, and one expressed discomfort with a paper-and-pencil test for measuring how well teachers assess student progress. The fourth scorer summed up their evaluation with the following statement:

> *I suppose that the most personal and concise way to convey my evaluation of this instrument is to say that I would not want my competency of assessment practices and skills assessed with it in its present form and given the accompanying scoring criteria, particularly without receiving the accompanying training.*

Assessment Format

Format Features

The format employed by the Assessment of Competence in Monitoring Student Achievement in the Classroom is that of a paper-and-pencil test with brief written scenarios serving as stimuli to which teachers respond in writing. The assessment also consisted of a design which involved testing before and after completion of a series of staff development workshops.

Clarity of Assessment

Because this assessment was in the developmental stage, the focus of the evaluation questions with respect to the assessment was on identifying problems in the assessment which could affect teacher responses. Teachers were only asked to elaborate on their negative responses; some teachers also elaborated on positive responses.

Clarity of questions. Teachers were asked if they believed that the directions for each exercise were clear. Eighty-one percent (17 of 21) of the teachers completing form A as a posttest and 57% (12 of 21) of those completing form B believed that the directions were clear. Those teachers who found some questions to be unclear wished for a definition of some specific terms, didn't understand the point of some exercises, or encountered directions which they believed contradicted their training, e.g., write a multiple-choice question assessing higher-order thinking skills when at one point the training stressed the that multiple-choice questions were not the best approach (although the training also included practice in writing such multiple-choice questions.)

The scorers were asked if there were any exercises with which the teachers consistently had difficulty. Their responses identified several exercises where they believed that improved directions would have oriented a number of teachers whose answers went astray. These exercises included:

- the exercises on both forms asking teachers to evaluate the soundness of various grading practices. Sometimes, the scoring criteria required two reasons to be given in order to receive full credit; however, this was not indicated in the directions, and many teachers gave a single reason.

- the exercises on both forms addressing cultural norms which interfere with assessment. This exercise was misunderstood by many of the teachers. Instead of addressing how different assessments could avoid conflicting with cultural norms, teachers talked about either how to counter stereotypes in the classroom or addressed psychological or interpersonal issues rather than cultural ones. In addition, many teacher responses indicated cultural insensitivity in that they believed that the solution to the problem was varying degrees of encouragement to produce the desired behaviors (e.g., if a student's culture discouraged participation in classroom discussion, some teachers suggested calling on the student more often).

- the exercise on one form focusing on defining features of a valid writing assessment. Instead of identifying features of a writing assessment (e.g., a pre/post design), teachers identified features of the writing to be evaluated (e.g., grammar or presence of a theme).

- the exercise on one form which asked teachers to compare three methods of assessing reading. Teachers commented more on the goals of reading instruction exemplified by these methods than on the assessment methods themselves.

- the exercises on both forms requiring the construction of multiple-choice items. Several teachers constructed fill-in-the-blank or matching questions instead of multiple-choice questions. However, the term "multiple-choice" did appear in capital letters in the instructions. The consensus among the scorers was that the directions specifying the type of answer to be provided should have been separated from the introductory information into a concluding paragraph, providing a focussed set of instructions for the teachers.

In addition, the passages about which teachers were to write multiple-choice questions were considered by several scorers to be problematic. Neither passage appeared to be written to deliberately stimulate higher order thinking, so construction of questions measuring higher-order thinking skills was challenging and required going well beyond the information presented in the text. One scorer summarized the problem: "I found the passage itself so poorly written and difficult to comprehend that I wasn't sure how anyone could write a reasonable test question that went beyond recall but that students could answer."

- In one exercise on one form, the term "draw the comparison" was used in the context of comparing two methods of assessing student understanding of differences between countries. Some teachers took this literally as meaning a physical drawing, while others assumed that the term meant a verbal comparison. (The latter was meant by the assessment developer.)

Teachers also were easily distracted by exercises which contained more than one issue, for example, a verbal exchange between a teacher and student concerning a cheating incident which asked the candidate to assess whether or not the teacher took an appropriate course of action. When the presenting dilemma was complex, teacher responses sometimes reflected goals other than assessment, e.g., penalizing inappropriate behavior in the cheating incident described. This may result from teachers discussing issues with which they feel more comfortable, which due to insufficient instruction and practice in assessment, is unlikely to be the measurement of student achievement. However, the presence of competing goals is precisely the context in which many assessment decisions will be made. One possible remedy for this is to recognize the competing goals in the presentation of the scenario, acknowledge that assessment goals have to be balanced against the competing goals, and ask the teacher to discuss the prompting problem solely in terms of accomplishing accurate assessment. For some of the exercises such as the cheating incident, this method of revision runs the danger of trivializing the problem, reducing the question to be answered to "Does this solution negatively affect the measurement of achievement?"

Clarity of scoring criteria. Scorers were asked if they "had any difficulties in applying the scoring criteria for any of the assessment exercises." Every scorer indicated that they had difficulty with some exercises, with two of the scorers describing difficulties with eight of the twenty exercises. General difficulties included:

- determining whether different statements are redundant or contain separate points
- judging when a multiple-choice item is testing recall vs. higher-order thinking skills, especially when teachers fail to follow instructions that ask them to label the latter items and the item is poorly written
- differentiating between general platitudes or instructional strategies not specifically related to the problem (which were not to be given credit) and responses which are specific enough to warrant credit
- judging ambiguous responses which combine discussion of methods of assessment with other issues, e.g., goals of reading
- determining whether a response is an example of a writing trait which is more than simply a variation of mechanics

Many of these problems could be solved through the provision of more examples during training that are more typical of those being scored, and by providing more precise definitions (either conceptual or through a series of contrasted examples) of the borderline between acceptable and unacceptable responses. Revision of the instructions for some exercises to provide candidates as well as scorers with an orientation toward the type of response expected would also facilitate scoring.

Evaluation of staff development training. Teachers who participated in the staff development training were asked to evaluate the sessions for two purposes: improving their performance on the posttest and for improving their ability to monitor student achievement in their own classroom. Twenty-nine teachers responded to the survey.

In terms of improving their performance on the posttest, 10 percent (3 teachers) evaluated it as "very useful"; 78% (22 teachers) rated it as "somewhat useful"; and 14% (4 teachers) rated it as "of little use." Several teachers praised the handouts and information presented. Many teachers reported being overwhelmed by the amount of information, citing a difficulty in recalling content presented in the early sessions, a need for more depth and less breadth in the material presented, and a lack of time to complete the reading of the handouts during the school year.

As for improving their ability to monitor student achievement in their own classrooms, 24% (7 teachers) of the teachers evaluated it as "very useful"; 59% (17 teachers) of the teachers rated it as "somewhat useful"; and 14% (4) of the teachers rated it as "of little use." One teacher rated it as somewhere between "very" and "somewhat" useful. Several of the teachers felt that it made them better critics of tests, both published tests and tests they developed themselves. Some teachers wished for more examples at the primary (K-3) level. Other teachers felt that it was too early in their teaching career to assimilate the information, and hoped that they could find ways to apply it through summer study of the handouts.

Over half (59% or 17) of the teachers reported using some of the principles presented in the staff development sessions in their monitoring of classroom achievement.

Cost Analysis

We will outline cost estimates for administering and scoring the Assessment of Competence in Monitoring Student Achievement in the Classroom based on the current status of the draft assessment that was pilot tested. We will also report the costs for developing this prototype and for pilot testing it. These costs should be taken as only preliminary estimates for costs that would be incurred if an assessment like this were to be further developed and modified for implementation on a wide scale.

Administration and Scoring Costs

This assessment is administered in a large group setting. The assessment can be administered by one or more persons with little or no training in the specific content of the assessment using procedures common to standardized group test administrations.

Scoring the assessment requires training raters knowledgeable in the content of the assessment. Scoring of the pilot test data, which included both training and actual scoring, required two days for six scorers. The majority of this time was devoted to scoring the assessments with only approximately one-half day devoted directly to training.

The six raters were able to complete nearly 400 ratings in the one and one-half days devoted primarily to rating. Dividing the 400 ratings by nine scorer days (i.e. six assessors times 1.5 days) results in an estimate that a scorer can rate approximately 40-50 assessments each day or approximately 4-6 per hour. Using $160/day as the cost for an scorer results in an estimate of approximately $3-4 per assessment rated.

Training costs can also be estimated using the pilot test scoring experience. The half-day training in the pilot test cost approximately $80/scorer plus the costs for the trainer. In the pilot test the trainer costs were distributed across only six scorers. It would be feasible to expand the number of scorers that could be trained in one session. Increasing the number trained to 10 would result in an estimate of $960 as the cost for a one-half day training for the 10 scorers, including an allocation of $160 for the trainer. If it is assumed that after training a scorer would spend two and one-half days rating and in this period could rate 40 assessment/day, the training costs would be distributed across 100 assessments per scorer or 1,000 assessments for 10 scorers. Thus, a half-day of training would cost approximately $1 per assessment. However, the scorers found a half-day training to be inadequate; recommended changes would result in training lasting at least one day, doubling the estimated cost to $2 per assessment. Combining the training and rating costs would result in an estimate of $5-6 per assessment for scoring this prototype assessment.

Costs for test administration, duplication of materials, postage, travel, etc. would also need to be added to the costs for scoring the assessments. We have used $30 per assessment for other large scale, group administered assessments as an estimate for these costs. Combining these results in the following estimated cost for administering and scoring this prototype assessment:

Training and Scoring:	$5-6 per assessment
Administration/Other Costs:	$30 per assessment
Total Administration and Scoring Costs	$35-36 per assessment

Development and Pilot Testing Costs

Costs for developing this prototype assessment were $83,971 and are broken out by general cost categories in Table 10.6 which also includes costs for pilot testing. These development costs are the expenses for the assessment developer to deliver the prototype assessment forms to the CTC and SDE. Additionally, $15,687 was expended in conducting the pilot test. These costs included those for FWL staff to observe teacher training provided

TABLE 10.6

DEVELOPMENTAL AND PILOT TEST COSTS FOR THE ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Cost Categories	Development	Pilot Testing
Staff-Salaries & Benefits	$42,632	$ 8,801
Consultants (Teachers, assessors, and other consultants)	10,000	1,840
Travel (Consultants and staff)	9,176	1,278
Other Direct Costs (Site rental, phone, duplication)	6,394	465
Total Direct Costs	$68,103	$12,384
Indirect Costs	15,868	3,303
Total Costs	$83,971	$15,687

by the assessment developer in which new teachers were trained in the concepts covered by the assessment.

These provide samples of developmental costs that should be considered if a similar assessment were to be adapted for use.

Technical Quality

Development

This assessment was developed as a successful response to a request for proposals to develop innovative forms of assessment for possible use in the licensure of new teachers in California. The assessment content was chosen based on a decade of research and development work by NWREL staff to identify task demands of classroom assessment and to design staff development activities which enable teachers to meet those demands. To develop the two forms of the instrument, NWREL staff constructed over thirty original exercises. Following review by members of the California Interagency Task Force, and then editorial revision, the exercises were assembled into pilot test packages and administered to a small number of teachers. Teachers participating in this "shakedown" were interviewed and asked to complete a questionnaire. Both the teachers' performances on the exercises and their opinions of the exercises were used to eliminate exercises which clearly did not work well. Exercises retained for further analysis were revised and edited to improve their clarity. After another review by the Interagency Task Force staff, the exercises were reviewed for potential bias, and were revised to eliminate any bias found.

The final phase of assessment development was the creation of scoring criteria for each exercise. Exercises for which clear and defensible criteria could not be identified were eliminated. The preliminary scoring criteria were tested against a small subset of the pilot test responses, which resulted in major revisions.

The resulting set of prototype exercises were then divided into two sets such that the content coverage was parallel. In many cases, this meant creating parallel forms of the same exercise. In other cases, different exercises covering the same material were chosen. An attempt was made to represent a variety of levels and subjects specified in the contextual information for each exercise. Another round of revisions based on further review by Interagency Task Force staff resulted in the two forms pilot tested in this assessment.

Reliability

The following analyses were performed on the pilot test data of the 50 teachers who took the pre-test and the 46 teachers completing the post-test. Interrater agreements were examined to assess the degree to which the scorers were able to consistently judge candidates using the scoring protocols provided. Internal consistency estimates were generated to assess the degree to which the variables or factors within each of the activities would form a measure and the degree to which the different activities related to each other and might form an overall assessment of a candidate.

Interrater agreements. The first measure of agreements among scorers was the differences in total scores between scorers rating the same candidate responses. These differences are presented in Figure 10.1. The differences were sorted into three groups: paired scores differing by 0-2 points, those differing by 3-5 points, and those differing by more than six points.

The degree of agreement among raters was such that only 15-18 percent of the rating pairs differed by 6 or more points which represents approximately 10 percent of the total possible. Ideally, it would be desirable to have a larger percentage of the rating pairs be within 0-2 points. But, given the draft nature of this prototype, these results suggest that the scoring criteria and system is such that raters can make similar judgments. The ability of the raters to achieve consensus on the ratings with these scoring criteria should also be interpreted in light of earlier comments by the raters about the scoring system. This is, although raters could understand and follow the rating criteria, they have also identified areas in which revisions should be considered.

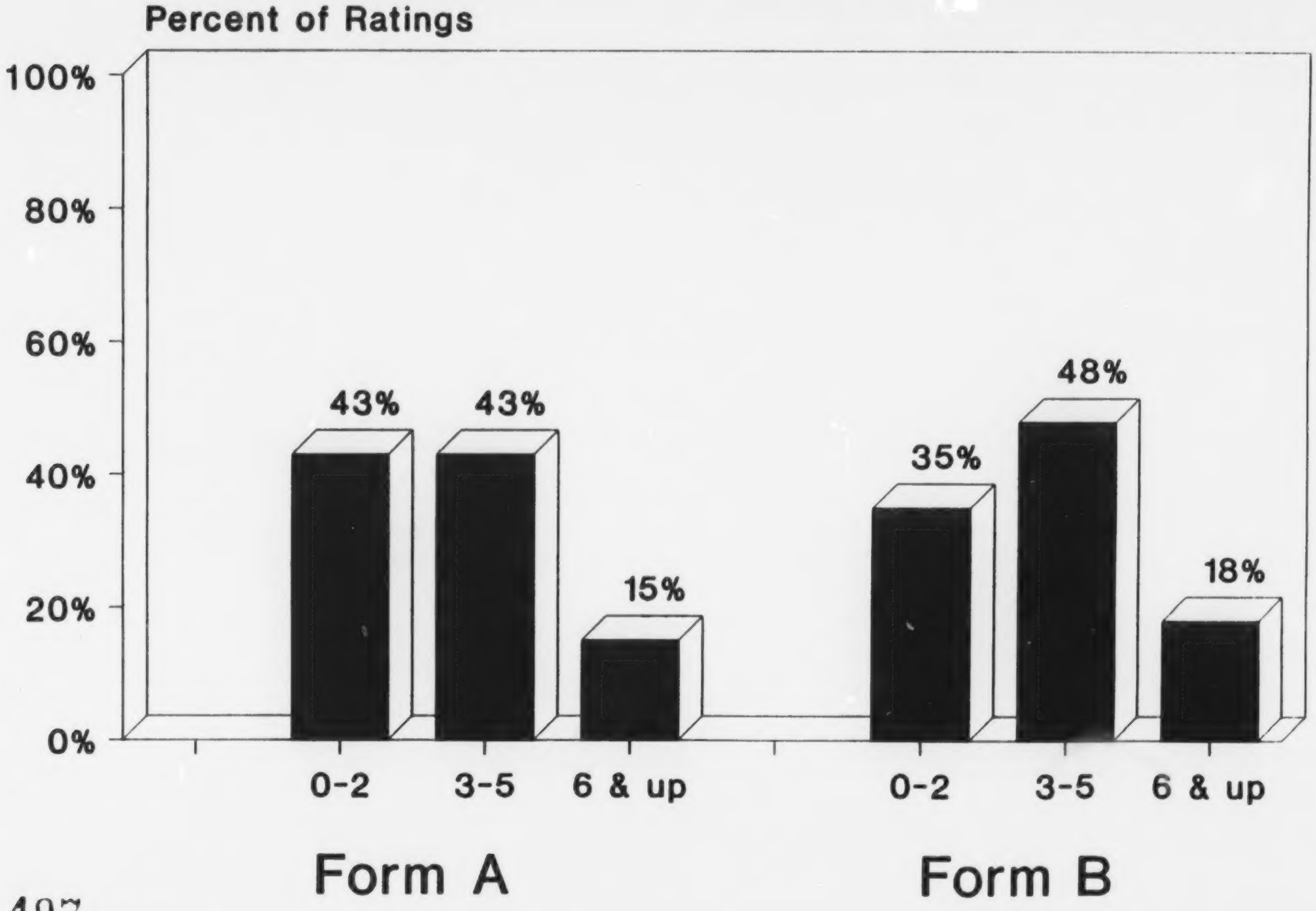
Interrater correlations. Correlations between scorers also serve as an estimate of interrater agreement. The correlations among rater pairs are displayed in Table 10.7. Correlations were calculated for total scores separately by form and rater pair. The average correlations across rater pairs were also calculated for each form.

These data also support a conclusion that agreement among raters can be achieved using the current criteria and system. It is likely that further refinements and revisions could result in even closer agreement between raters.

Internal consistency of the assessment forms. Coefficient Alpha reliability estimates were calculated for the two forms by using the individual ratings for exercises or their

FIGURE 10.1

Rating Differences Between Scorers



10.41

TABLE 10.7

CORRELATIONS BETWEEN PAIRED RATINGS FOR THE ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Form	Rater Pair*					Averaged Pair Ratings
	1	2	3	4	5	
A	.86 (15)	.88 (16)	.84 (16)			.86
B	.88 (6)	.80 (5)	.89 (5)	.67 (16)	.87 (16)	.84

*Rater Pairs for Form A and B are different. The numbers in parentheses are the number of teachers scored by each rater pair.

subparts. Calculations were also done separately by whether or not the form was a pretest or a posttest. These Alpha reliabilities are listed below:

Form	Reliability
Form A (N=48)	.67
Pretest (N=26)	.70
Posttest (N=22)	.68
Form B (N=48)	.53
Pretest (N=24)	.56
Posttest (N=24)	.52

In contrast to the agreement among raters, the degree of internal consistency evidenced within the prototype assessment forms is modest to low. This suggests that in its current form the assessment does not form an overall measure of a single factor of teachers' knowledge of monitoring classroom practice, but includes items that measure somewhat independent factors. Further development would be needed to determine whether the assessment in this area would result in multiple factors and measures or whether the assessment might form a more homogenous measure of teachers' knowledge in this area.

Validity of Agreement through Group Comparisons

Differences in performances were examined for male/female, primary/intermediate grade, inner city/non-inner city, and white/minority teachers. This section uses the pilot test data to look for indications of any possible group differences in performance on the assessment. The pilot test sample size and design were not constructed to provide information sufficient to provide stable estimates comparing differences among these groups.

For instance, there were only two minority teachers in some of the groups. Nevertheless, an examination of differences among groups provides some initial insights into the validity of this assessment. Table 9.8 contains a summary of trends for the pilot sample of 42 teachers who completed evaluation surveys, including demographic information. Appendix G provides the means, standard deviations and numbers of candidates from which these summaries were constructed. A plus (+) indicates that the mean or average for the first group was greater than that for the second group. For example, the pluses for the first column indicate that for two of the four tests, the average female score was greater than the that of the males.

Trends in the table indicate that for this pilot test sample:

- the average score of females was higher than that of males half the time;
- the average score of primary (K-3) teachers was higher than that of intermediate (4-6) teachers on three of the four tests;
- the average score of inner city teachers was lower than those of non-inner city teachers on all four tests;
- the average score of white teachers was higher than those of minority teachers on three of the four tests.

If these trends were to hold for larger, more representative samples, the only encouraging trend would be that for gender. No differences are large, and the average male score is higher than the female no more often than the reverse. The small sample sizes preclude the drawing of any conclusions about how the different groups of teachers might perform on this type of assessment. However, these trends suggest that any further work on assessments of this type should include the examination of possible differences between primary and intermediate teachers, inner city and non-inner city teachers, and white and minority teachers.

Content validity. Evidence of the content validity of this assessment comes primarily from two sources. The first is the decade-long experience of the assessment developer in developing a curriculum for teaching assessment of student competence in the classroom, based not only on research but also on feedback from practicing teachers. The second is

TABLE 10.8

TRENDS OF MEAN DIFFERENCES IN PERFORMANCE BETWEEN CANDIDATES WITH DIFFERENT CHARACTERISTICS*

ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT IN THE CLASSROOM

Form	Gender Female/ Male	Level of Teaching K-3/4-6	Teaching Location Inner-City/ Non-Inner City	Ethnicity Non-Minority/ Minority
A				
Pretest	+	+	-	+
Posttest	-	-	-	+
B				
Pretest	+	+	-	+
Posttest	-	+	-	-
SUMMARY	2/4	3/4	0/4	3/4

*Entries reflect the direction of the mean differences for the different candidates. For example, on the pretest for Form A, the average mean of female teachers in the pilot test was greater than the males. These do not generally represent statistically significant differences and due to small N's no tests of significance were calculated.

the analyses of the congruence between the assessment and the various Model Curriculum Guides and the California Beginning Teacher Standards.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Assessment of Competence of Monitoring Student Achievement in the Classroom, organized into the areas of administration, scoring, content, format, and a brief summary.

Administration of Assessment

Like other large-scale examinations, the Assessment of Competence in Monitoring Student Achievement in the Classroom is administered simultaneously to a large number of people. Benefitting from many years' experience in conducting such examinations, the administration of the actual assessment poses few logistical problems. The most crucial logistical activity is the selection of the assessment and staff development sites. Although the assessment portion of the Assessment of Competence in Monitoring Student Achievement in the Classroom is less expensive to administer than many of the other assessments piloted, the economy of scale achieved depends on the number of teachers participating at a single site. Therefore, the higher degree of centralization afforded by an assessment of this type may place larger burdens on teachers from rural areas who will have to travel some distance to a selected site.

The exercises varied in their susceptibility to memorization of standardized answers which allow a teacher to pass the assessment without the knowledge and ability to apply the principles tested. The exercises which appeared to be least vulnerable to this type of coaching were more performance-based, and teacher responses were very dependent on the subject-matter content presented in the stimulus exercise. The exercises which were judged to be highly vulnerable were those involving listing of general principles which would be invariant across subject matter. We recommend that any future development of instruments of this type focus on the more performance-based exercises.

Scoring

The scoring system is in need of further developmental work. The system for training scorers could benefit from the following improvements:

- providing opportunities for scorers to practice scoring independently, with the scoring compared against a standard and problems in scoring discussed as a group;
- a greater number of examples of scored responses;
- examples that more closely resemble those to be scored; and
- more precise definitions (either conceptual or through a series of contrasted examples) of acceptable and unacceptable responses.

These revisions to the current scoring system should reduce the number of problems reported with its implementation and provide scorers with more concrete guidance to evaluate teacher responses.

Assessment Content

Our observations and information collected from scorers and teachers participating in the pilot test suggest the following conclusions about content:

- Modifications of the exercises are necessary to bring the assessment into closer congruence with current curriculum guides and frameworks.
- Most, if not all, of the assessment methods which might be used to assess the more in-depth knowledge called for in the latest curriculum guides and frameworks are represented in the current collection of exercises.
- The exercises vary in their ability to evaluate a teacher's competence with respect to student diagnosis, achievement, and evaluation, as set forth in the California Beginning Teacher Standards. The more promising exercises were performance-based.
- Most of the teachers believed that the exercises had relevance to their task of monitoring student achievement.
- Teachers were mixed in their opinions as to whether they had been sufficiently prepared for the assessment. This was true of the teachers who participated in the staff development activities as well. Those who did not attend the staff

development workshops attributed their insufficient preparation to a lack of relevant coursework, while those completing the training cited a need for practical application and/or more experience.

- As is true of many of the assessments pilot tested to date, the content of the Assessment of Competence in Monitoring Student Achievement in the Classroom focussed on the higher grade levels covered by the credential. Teachers were split on the appropriateness of the assessment for teachers at different grade levels, with half believing that it was fair across different grade levels, and half believing that it was not.

- The assessment was designed to portray students from a variety of cultural backgrounds, as well as men and women in nontraditional roles.

- Teachers overwhelmingly believed that the assessment was appropriate for teachers of diverse student groups. An expert on teaching diverse students expressed concern about several aspects of the assessment. The first is the ability of the assessment to accommodate alternative conceptions of teaching, particularly those which might be particular to specific types of classrooms. The assessment, for example, does not account for the contextual assumptions that are implicitly made by candidates and scorers. This is particularly important because little contextual information is given in each exercise. Finally, the scoring guide do not provide for the possibility of correct responses which were not previously identified.

- Teachers overwhelmingly believed that the assessment was fair to different groups of teachers. Scorers had mixed views, with concern expressed about (1) the possible advantage of people with strong written communication skills over those whose strength lay in oral communication skills, and (2) the possibility of passing people who had mastery of the technical language and abstract principles but not mastery of their application.

- Less than half of the teachers believed that the assessment was an appropriate way of assessing their competency in evaluating student achievement, preferring instead methods which examined their actual assessment of their own students.

Assessment Format

The format is that of a paper-and-pencil test with brief written scenarios serving as stimuli to which teachers respond in writing, accompanied by a series of staff development workshops.

Based on evaluations by teachers, scorers, and FWL staff, the following modifications in the assessment instruments are needed:

- Revision of the directions for the exercises so that they clearly indicate the scope of the expected answer.
- Identification of those exercises where a singular focus on assessment is likely to have a negative impact on other goals (e.g., self-esteem, classroom management). These exercises should either be eliminated or revised. Revisions should include acknowledgement (1) of the competing goals; and (2) that considering the effects on assessment is only one step in evaluating the practice portrayed.
- Elimination of the goal of designing "self-evident" scoring criteria, along with a corresponding expansion of the training of scorers.

Most teachers rated the staff development training as "somewhat useful" in terms of improving their ability to monitor student achievement in their own classroom; however, compared to teachers who did not participate in the staff development workshops, the staff development workshops did not improve their assessment scores noticeably.

Summary

The Assessment of Competence in Monitoring Student Achievement in the Classroom needs substantial revisions in prompt materials, scorer training, and scoring criteria before it could be considered for use in licensing in California. Such revisions should proceed toward the development of more performance-based exercises, where teachers are asked to actually develop a portion of an assessment.

CHAPTER 11:

CONCLUSIONS

This final chapter contains our conclusions from our pilot testing experience during the second year of the California New Teacher Project. The first section describes each assessment approach that was pilot tested during 1990-91, and identifies strengths and weaknesses. Next, cost estimates are discussed and conclusions about characteristics of successful training of assessors and scorers are stated. The chapter concludes with a discussion of policy issues that have been identified during this round of pilot testing.

Assessment Approaches

Although the purpose of the pilot tests was to use the specific instruments to learn about the potential of assessment *approaches*, the preceding chapters focused on the individual instruments. This section describes the assessment approach for which each instrument serves as an exemplar, and summarizes our conclusions about the critical features as well as the strengths and weaknesses of each assessment approach. These conclusions are tentative for some assessment approaches, as few of the instruments piloted represent an assessment approach with a lengthy history with respect to teaching.

Each instrument reflects one of four assessment approaches: structured simulation tasks, classroom observation, videotaped teaching episodes, and a set of performance-based assessment center exercises.

Structured Simulation Tasks

Definition. This assessment approach, administered in a large group-setting, requires a teacher to perform a task which simulates work characteristic of one or more teaching responsibilities. The teacher's response is then compared to a list of previously identified responses or response characteristics.

Characteristics of instruments piloted. Three instruments representing a Structured Simulation Task approach were piloted during the second year of the project. Two of them, the Secondary Life/General Science Teacher Assessment and the Structured Simulation Tasks for Secondary English Teachers, are modeled after an assessment created by the Rand Corporation for use in the licensing of lawyers. The other, the Assessment of

Competence in Monitoring Student Achievement in the Classroom, is based on a decade of experience in staff development for teachers in the area of student achievement.

All three instruments ask teachers to do one of the following: to analyze a completed teaching task, to actually perform a brief task, or to outline how they would perform a larger task. The exercises which form the Assessment of Competence in Monitoring Student Achievement in the Classroom focus on either small pieces of a larger task (e.g., construct three items for a multiple-choice test) or on general outlines of how larger tasks (e.g., construction of an observation assessment to determine readiness for promotion to first grade) could be accomplished. Answers for this instrument are relatively brief, typically answered in a sentence or two or by listing up to four items. The tasks piloted for the other two assessments focus on large tasks such as critiquing a lesson or planning a two-week unit. All but one task are divided into subparts. Background material which describes the classroom context is provided for each task. This requires the teacher to take more factors into account in developing a response, which takes such forms as a list that contains up to twenty items, written responses to samples of student writing, or an outline of a unit plan.

The scoring systems for the instruments differ somewhat. Both the Assessment of Competence in Monitoring Student Achievement in the Classroom and the Secondary Life/General Science Teacher Assessment used analytic scoring systems which compared teacher responses to a predetermined set of response characteristics corresponding to varying numbers of points. Scorers for the Secondary Life/General Science Teacher Assessment were allowed to use their discretion in allowing credit for answers not on the scoring key, while scorers for the Assessment of Competence in Monitoring Student Achievement in the Classroom were asked to only note, but not credit, possible additions to the scoring key. While the scoring system for the Structured Simulation Tasks for Secondary English Teachers began with an analytic design, many of the analytic scoring guides were abandoned during training in favor of a holistic scoring approach. Although each assessment contained pieces that were similarly scored, the nature of the majority of the scoring criteria for the English and science assessments differed from that of the evaluation assessment. The scoring criteria for the Secondary Life/General Teacher Assessment and the Structured Simulation Tasks for Secondary English Teachers generally focussed on characteristics of a product that the teacher produced which were entirely dependent on the task and the subject-matter content. In contrast, the scoring criteria for the Assessment of Competence in Monitoring Student Achievement in the Classroom most often focussed on teacher criticism of general assessment practices or on general principles of construction of particular types of student assessments which were content-free.

Strengths and weaknesses. The major strengths of the Structured Simulation Tasks approach to teacher assessment are: (1) ease of administration and scoring; (2) job relevance through a focus on application, especially in the area of content pedagogy; and (3) the ability to assess teaching of diverse students through use of descriptions of specific types of students in the stimulus materials. Structured simulation tasks can be easily administered on a large scale, and do not require administrators with content expertise. The job relevance of specific components varied, but was strong for those tasks or exercises in the pilot tests which asked the teachers to produce or analyze some sort of product related to instruction (e.g., lesson description, multiple choice items), analyze a simulated classroom transcript illustrating effective and ineffective instructional techniques, or perform a task related to laboratory safety. These tasks give a clear idea of whether or not the teacher can produce acceptable work in the context described or whether the teacher can analyze teaching products or interactions of other teachers. Since beginning teachers typically create only a small portion of their instructional materials, the ability to analyze activities and materials is important. Specifying the type of student and/or instructional goals in the stimulus materials is also important for the proper design of instruction and for ascertaining a teacher's ability to design instruction for various types of students.

The major weaknesses of this approach are (1) the paucity of diagnostic information generated by the scoring system; (2) the ability to reflect only a few specified teaching contexts, techniques, and topics in the stimulus materials without vastly increasing the time for administration; and (3) the difficulty or possible inability to measure many teaching competencies involving either nonverbal behavior (e.g., some aspects of classroom management, establishment of rapport) or the classroom as a whole (e.g., efficient management of routine activities such as collecting homework).

While the scoring system indicates whether or not a teacher can perform the task in the context described with the teaching techniques described, it cannot provide diagnostic information as to the source of the teacher's difficulties, i.e., did the teacher fail because of a lack of knowledge of the content pedagogy, a lack of knowledge of the specific group of students in the stimulus materials, or a lack of experience with the specific techniques cited? Thus, while this approach may be suited to licensure decisions, it would be unlikely to provide sufficient information to guide choice of staff development activities.

This methodology also assumes that either beginning teachers have the ability to analyze situations which do not resemble their own or that enough situations are represented that the teacher is not unduly penalized due to his or her lack of breadth of experience. It is not clear whether or not this assumption is warranted. Tailoring responses to the type of students, teaching techniques, and/or topic specified in the tasks was difficult for teachers, based on their responses and feedback evaluations. It is possible that these difficulties might be eliminated through improved and expanded instruction during teacher preparation. On the other hand, it is equally possible that the difficulties are characteristic of a beginning teacher who has limited experience, and that the ability to generalize to different students, different teaching techniques, and different topics does not fully develop until a later date. Choosing between these two alternative explanations is difficult until more information is derived from the current research on teacher preparation and differences between beginning and experienced teachers.

Some teaching skills are difficult to simulate, e.g., the establishment and maintenance of rapport between teacher and students, and efficient management of routines, and thus these skills are probably better assessed with other assessment approaches.

Other possible weaknesses of the approach are in the areas of fairness across groups of teachers and the appropriateness across different teaching contexts. Given the limited diversity in our sample of teachers and the lack of alternative measures of their teaching skills, this issue needs further exploration before drawing any conclusions. However, our consultant on diversity warned that the sample task that she examined had the potential to penalize specific groups of teachers for teaching behaviors which were effective in their specific context. (The specific example cited, with a citation of research attesting to its effectiveness, was the use of sarcasm by black teachers to motivate black inner city students.) She expressed concern that an approach that only focuses on one method of teaching or on one context might not be appropriate for these teachers who are effective in their specific context and who evaluate teaching techniques in light of their experience in that context.

Classroom Observations (Subject-Matter Focus)

Definition. A classroom observation approach to teacher assessment consists of observing teachers as they instruct students in their classrooms. This approach was reviewed in the Year One Report (Estes et al., 1990). A classroom observation with a subject-matter focus includes specific categories which examine the effectiveness of instruction in a particular subject.

Characteristics of instrument piloted. The Science Laboratory Assessment instrument piloted is a closed system, high-inference instrument. It requires observers to use their professional expertise to make judgments about specific categories of teacher behaviors. The innovative aspect of the Science Laboratory Assessment is the inclusion of several categories specifically designed to assess the subject-matter pedagogy and safety skills of laboratory science teachers at both the elementary and secondary levels. The scoring system by which this subject-specific evidence was evaluated, however, only reached a preliminary stage of development, and needs much further refinement before the observation instrument could be successfully implemented.

Strengths and weaknesses. The major strength of all classroom observations is job relevance. Classroom observations assess teachers in the process of doing their work, so they have high job relevance and face validity. When teachers name a method of preferred assessment, they usually name classroom observations. In order to achieve this strength, however, classroom observations, whether subject-specific or not, need clearly established foci for observations, and criteria for assessing the adequacy of the teacher performance observed. In addition, the observers must be trained to recognize similar phenomena as they occur in quite different contexts.

A strength of classroom observations with a subject matter focus is that such observations allow assessment of some subject-specific teaching competencies which are difficult to assess except through direct observation, such as the maintenance of a safe environment for laboratory science teachers, and initiating and managing discussions among students in different subject areas.

The weaknesses inherent in all classroom observations are (1) the lack of generalizability across teaching contexts and topics, and (2) the complexity of administration and observer training. Classroom observations have limited ability to sample, as observations are limited to a specific classroom of students, a specific lesson, and a specific time of year. Stable estimates of teaching competencies depend on multiple observations, and are not generalizable across lesson types, subjects and grade levels (Stodolsky, 1988, p. 12).

Although all classroom observations involve some administrative complexity, observations with a subject matter focus increase the complexity. For a valid assessment, a careful match must be made between observers and teachers with respect to grade level (e.g., elementary, secondary), subject matter (e.g., life science, physical science), and availability. Enabling observers to make comparable judgments while watching instruction in differing teaching contexts and on different topics requires complex and lengthy training if high-inference observation instruments such as the two pilot tested are used.

One area of difficulty experienced in the Spring pilot test may be a symptom of an additional limitation of high inference classroom observations in general, but particularly those observations like the Science Laboratory Assessment which attempt to assess both general pedagogical skills and subject matter skills. The Science Laboratory Assessment, a subject-specific instrument, requires the observers to rate more domains than the Connecticut Competency Instrument (CCI), an instrument which focuses on general pedagogy. Observers using the former instrument seemed to have some difficulty not experienced by observers using the latter instrument in gathering sufficient evidence to support judgments for each domain. Some of this difficulty could probably be addressed by further refining the training for gathering evidence and with additional experience in administering the instrument. However, it is possible that there are limits to the number of domains that can be assessed through a single observation using instruments like the two piloted, which are extremely high-inference in nature.

Semi-Structured Interviews

Definition. Semi-structured interviews provide opportunities for candidates to respond orally to a standardized series of questions or tasks that are presented verbally by an examiner who uses a script known as an interview schedule. Semi-structured interviews include "probes" to be used at the administrator's discretion to enable teachers to elaborate on their responses.

Characteristics of instrument piloted. Similar to the other semi-structured interviews piloted, the SSI-SSS asks teachers to perform a teaching task, and then respond to questions about their decisions. The questions often include a set asking how their decisions would change if they were teaching a homogeneous class with respect to ability, exploring adaptations for both high ability and low ability classes. Four tasks related to a common instructional topic constituted a complete interview. Although interviewers used a structured interview protocol, they were instructed to ask additional questions probing a teacher's answer if the answer was considered to be unclear or ambiguous. The interview was recorded on videotape for later scoring.

Strengths and weaknesses. The major strength of the Semi-Structured Interview assessment approach is its ability to measure the depth of a teacher's knowledge about how to perform specific instructional tasks. This measurement of depth depends on carefully constructed questions that elicit the thinking that lies behind the decisions made. Interviews are especially effective in facilitating the display of knowledge that impacts instructional planning and decision-making. Questions explicitly directed to how student characteristics affect instructional decisions can also give a sense of the range of student characteristics that are taken into account in instructional decisions as well as the extent to which instruction is tailored to students.

The major weaknesses of Semi-Structured Interviews are: (1) the ability to reflect only a few specified teaching topics in the stimulus materials without vastly increasing the time for administration, and (2) the unexplored relationship of teacher descriptions of decisions to the application of those decisions in the classroom.

The measurement of the depth of a teacher's knowledge is accomplished at the sacrifice of the measurement of breadth. Because the interview is time-consuming, only a small number of topics can serve as the focus. Furthermore, the degree of a teacher's familiarity with the focal topic affects his or her depth of knowledge, so performances on interviews with different topics may not necessarily be highly correlated.

The interview features the teacher talking with another adult about instructional plans and decisions. The discussion focusses on hypothetical plans. Whether a teacher can actually apply those plans in a classroom while simultaneously attending to classroom management and unexpected student responses has not been measured to date. The interview performance may be strongly affected by a teacher's ability to improvise dialogue, which is unrelated or weakly related to their teaching ability.

Videotaped Teaching Episodes

Definition. This approach to teacher assessment requires a teacher to respond to questions pertaining to videotaped scenarios of teachers instructing students in a variety of contexts. Some supplementary material (in this case, stories read by students) may be provided.

Characteristics of instrument piloted. One instrument representing the Videotaped Teaching Episodes Approach, the Language Arts Pedagogical Knowledge Assessment (LAPKA), was pilot tested. This instrument centers around videotaped scenarios which vary in type of elementary language arts instruction, grade level, and group size. Each scenario is broken down into short segments. After previewing the questions for a particular segment, teachers view that segment once and respond to the questions with short-answer written responses. The questions ask the teachers to describe important features of the content pedagogy represented in the videotape, evaluate the effectiveness of these methods, and extend the principles inherent in the methods to suggest ways of improving or changing the methods shown.

The structure of presentation and response used in LAPKA contrasts with other possible variations within this assessment approach. For example, the Stanford Teacher Assessment Project piloted four assessments in four different subjects (elementary literacy, elementary mathematics, secondary biology and secondary history) using a videotape stimulus to identify master teachers. In this variation on the Videotaped Teaching Episodes approach, teachers viewed a brief set of videotaped lesson segments, and then responded in a semi-structured interview to the material they viewed.

LAPKA's use of a single monitor differed from yet another variant on the Videotaped Teaching Episodes approach used by David Berliner (1989), where teachers ranging in experience viewed three monitors showing the same lesson from different camera angles and commented on what they saw. Berliner's experience suggests that the LAPKA approach was more suitable for beginning teachers, as his beginning teachers experienced difficulties in observing the multiple monitors simultaneously, while the experienced teachers were able to effectively use all three to interpret events. The LAPKA format of having teachers respond to specific questions which are previewed before seeing each videotaped segment is probably also helpful, as Berliner's beginning teachers had difficulties in focusing their attention during viewing of the videotape, particularly in distinguishing typical from atypical events and important from unimportant information.

Strengths and weaknesses. The strengths and weaknesses of the Videotaped Teaching Episodes approach are more difficult to identify than those of other assessment approaches because the assessment approach is relatively new, and a scoring approach that fully capitalizes on the use of the medium has yet to be developed. Our identification of strengths and weaknesses is, therefore, tentative.

The strengths of the Videotaped Teaching Episodes approach are: (1) job relevance through the actual portrayal of teachers in action; and (2) the ability to assess specific teaching knowledge which is difficult to assess using other assessment approaches. Teachers are asked to describe and/or evaluate a series of videotaped segments showing teachers instructing their students, as opposed to a stimulus of written outlines of lessons or simulated transcripts. The videotape stimulus is especially good for evaluating some aspects of knowledge of teaching, e.g., the ability to know when a student's nonverbal responses indicate that s/he is becoming too frustrated, and the ability to analyze teacher-student interactions. While assessing these abilities is a particular strength of Videotaped Teaching

Episodes, it is important to remember that there is much information possessed by the teacher in the videotape (i.e., knowledge of individual students and school context) that cannot be fully communicated to the teacher being assessed.

The major weaknesses of the Videotaped Teaching Episodes are (1) the expense and complexity of development associated with videotaping lessons; (2) the difficulty in assessing certain teaching competencies; and (3) a dependence on technology for administration. Development of suitable videotapes is a complex process and can be very expensive. The film and videotaping equipment and the extensive editing required to produce high quality videotapes contribute to the expense. The production of suitable lesson segments is a complicated process. Scripting can produce artificiality, while naturalistic videotaping may not produce results which lend themselves to assessment.

Due to technical difficulties in sound and field of vision, the videotaping is best suited for small groups of students, and not entire classrooms. Therefore, such competencies involving the classroom as a whole, such as maintenance of behavior standards and keeping students engaged, are difficult to evaluate except with multiple monitors, which Berliner (1989) finds are unsuitable for beginning teachers.

The administration of the assessment is moderately complex, due to the reliance on technology. The assessment rooms must be set up to ensure that each teacher being assessed has an equally clear view of the monitor, and that the sound is audible to all teachers. In addition, equipment failure, though likely to be rare, has severe consequences for the assessment, either extending or canceling the assessment administration. Equipment failure can be minimized through pretesting equipment and making arrangements for backup equipment and a technician.

Performance-Based Assessment Center Exercises

Definition. Performance-Based Assessment Center Exercises have two main characteristics: (1) they bring teachers together at a central place to participate in a series of activities, each of which uses a different methodology to measure a distinct teaching skill; and (2) the activities require the teachers to directly demonstrate some skill which can be assessed by evaluating either the performance or the product produced, depending on the focus of the activity.

Characteristics of instrument piloted. One instrument representing the Performance-Based Assessment Center Exercises approach was pilot tested: the Secondary

English Assessment. This instrument consists of three activities. The first activity asks teachers to respond to two samples of student writing to demonstrate their ability to analyze student writing and to communicate their analysis of the writing both to the student and to peers. The second activity uses a small group discussion format to measure a teacher's ability to analyze a text and participate in a group discussion. The third activity asks teachers to deliver an extemporaneous speech on a given issue pertaining to the use of language in the classroom to measure a teacher's speaking ability with respect to important issues in English instruction. Each activity was scored using a holistic scoring process.

Strengths and weaknesses. The major strengths of the Performance-Based Assessment Center Exercises are (1) job relevance through a direct focus on specific teaching abilities; (2) the in-depth measurement of a small number of distinct teaching abilities; (3) the possibility of multiple measures of a single teaching competency using different methodologies; and (4) the ability to assess the teaching of specific groups of students specified in the stimulus materials.

Although job relevance is a major strength of Performance-Based Assessment Center Exercises, the realization of this strength depends on the ability to simulate the skills being measured outside the classroom. For instance, in the instrument pilot tested, the Secondary English Assessment, the activity asking a teacher to respond to student writing samples is very similar to what a teacher does in the classroom. In contrast, the activity where teachers discuss a piece of literature measures a teacher's abilities to interpret a text and to participate in a group discussion about the text. While these abilities are important skills that English teachers should have, a more relevant job skill would be the ability to *teach students* to interpret and discuss literature.

Performance-Based Assessment Center Exercises focus on only a few specific teaching abilities, but measure multiple aspects of these abilities. This approach thus has the potential to provide rich diagnostic information for the few abilities assessed. In addition to assessing abilities, it has the potential to measure other more general domains of knowledge of teaching, such as knowledge of students, through several different methodologies, decreasing the chances of mismeasurement due to deficiencies in skills relating more to a specific methodology (e.g., verbal fluency with adults, in the case of Semi-Structured Interviews) rather than to the knowledge being assessed.

Like Structured Simulation Tasks, the Performance-Based Assessment Center Exercises assessment approach facilitates the assessment of the ability to teach specific types of students described or reflected in the stimulus materials. For instance, all of the stimulus materials for the Secondary English Assessment -- i.e., the samples of student writing, the literature discussed, and the topic of the extemporaneous speech -- reflect the teaching of students from diverse cultural and linguistic groups which is typical of the majority of California classrooms.

The weakness of the Performance-Based Assessment Center Exercises approach lies in three areas: (1) the inability to assess teaching competencies which are difficult to simulate, (e.g., a teacher's rapport with students or the establishment of classroom routines); (2) the small number of teachers that is assessed per assessor, when simulations using small groups or individual presentations are used; and (3) the complexity of scheduling candidates when only a limited number can be assessed through one or more of the activities in the set of exercises.

Portfolio

Definition. A portfolio is the documentation of actual teaching experience, through examples of what the teacher considers to be his/her superior work and materials related to an actual unit taught. Possible portfolio entries include lesson plans, handouts, student work with teacher responses, and a journal or self-reflective essay.

Characteristics of instrument piloted. The piloted portfolio asked teachers to document a three- to six-week unit of instruction in which the classroom activities are unified by a single focus (e.g., a novel, a particular genre, a set of skills). Required portfolio entries include an outline of the unit plan, a weekly log documenting the teaching of the unit, all materials and assignments given to students, samples of student work with teacher responses, student evaluations of the unit (or of one major activity), and an essay reflecting upon the teaching of the unit and lessons learned. Using a holistic scoring process, each portfolio was scored in six areas: planning abilities, unit design, portfolio presentation, general pedagogical abilities, subject-specific pedagogical abilities, and reflective ability.

Strengths and weaknesses. A portfolio assessment's major strengths are: (1) job relevance through the focus on artifacts and activities that occur during and after instruction of actual students; and (2) the depth of measurement of skills in the domains of planning and designing instruction.

Portfolios are a combination of (1) documentation used in or resulting from teaching and (2) teacher reflections on their teaching of their own students. These provide a rich source of data directly derived from a teacher's job activities. If well constructed and detailed, feedback from a portfolio assessment should be directly applicable to a teacher's work.

A portfolio provides an excellent opportunity to focus on the teaching skills of planning and designing instruction, as well as the ability to reflect on one's own teaching. However, as reflected in the performance results, planning and designing instruction appear to be difficult skills for beginning teachers. Therefore, feedback on the evaluation of these skills would be extremely beneficial, and potentially inform staff development efforts.

A portfolio assessment's major weaknesses are: (1) the length of time required to administer the assessment; (2) the complexity of scoring; and (3) its ability to capture oral activities in a classroom.

A portfolio is the assessment requiring the longest time for a teacher to complete. If it is desired to facilitate the process for beginning teachers, a contact person to answer questions should be available during the time required for completion. However, this increases the personnel time devoted to administration. In addition, some teachers in our sample experienced difficulty in completing the portfolio over the course of a designated semester. One teacher was ill for an extended period; another was reassigned to a different set of classes in the middle of the unit. Because the portfolio covers weeks instead of days, the probability of a disruption are higher than for other assessment approaches.

The complexity of portfolio scoring is minly attributable to the fact that portfolios address a wide variety of topics, so there is no standardized stimulus. Portfolios can focus on different strands within English, e.g., literature, writing, grammar, drama. Evaluation of a portfolio requires complex training so that scorers can make standardized judgements across portfolios.

Although videotaped entries were allowed, participating teachers chose to submit only written entries. Some teachers did not believe that the written entries portraying responses to student work constituted a fair representation of their skills in diagnosis and evaluation, because much of their efforts were oral. While this could be overcome by inclusion of a videotape portraying oral comments, viewing and evaluating the videotape would increase the complexity of the scoring.

Guidelines for the Design of Training

During the first two years of pilot testing, we have observed a variety of training sessions for assessors, observers, and scorers. In reflecting over the strengths and weaknesses observed in all these sessions, we have identified some guidelines for the design of effective training. These guidelines will not seem profound, especially to educators, since they are simply principles of good instruction. They may even seem trivial or obvious, but many of the problems we have observed in training can be traced to the failure to follow one or more of these guidelines.

First, there should be clearly specified performance criteria for trainees in any role in assessment administration or scoring, in terms of what the trainees should be expected to be able to do upon completion of the training. Examples of such criteria are:

- For assessor training, the ability to identify instances where teacher responses to questions require a followup question for clarity;
- For observer training, recognizing the same phenomenon in different teaching contexts or lesson types; and
- For scorer training, the ability to accurately match teacher responses to predetermined correct responses, even though the teacher responses might be phrased in completely different ways.

As with scoring criteria, these performance criteria may be more general in the early stages of an assessment instrument's development, but the lack of specific criteria generally means that the assessment is not clearly conceived. An erroneous or even too-strict specification of criteria is preferable to criteria which are too general. Inappropriate criteria become obvious when they are applied, and the nature of the dissatisfaction frequently points to more appropriate criteria; however, this guidance is not available when criteria are vague. Typically, this process of applying and then revising criteria seems to take at least two or three iterations.

The training itself should be focused on instructing observers, assessors, or scorers to develop the skills needed to meet these performance criteria. In order for it to do so effectively, the following are necessary:

- Clear definition of all terms and criteria, with specific examples, to clearly communicate expectations and standards.

- Sample responses spanning the range of likely occurrences to illustrate judgments to be made, such as conditions requiring intervention by an assessor or the application of scoring criteria.

- Opportunities for the trainees to voice their opinions about good teaching with respect to the judgments they will be making so that the trainers can compare these perceptions against assumptions built into the assessment instrument. Sometimes, the trainees need to be guided to broaden their conception of good teaching or to redefine their standards to allow for fair application across different teaching contexts.

- Provision for independent practice in applying the skills being learned, together with the monitoring of individual performances and subsequent adjusting of instruction, if necessary.

- Provision for individual formative feedback to each trainee on their performance and a summative assessment at the end.

These guidelines are based on strengths and weaknesses observed in training sessions for the two years of pilot tests, and provide a framework for planning and evaluating training for various tasks in administering and scoring assessments.

Cost Estimates

Most of these assessment approaches, particularly those yielding diagnostic information which might inform staff development choices, are considerably more expensive than multiple choice tests. Our estimates of the per teacher cost of administering and scoring assessments such as these pilot tested range from $36 per teacher for the Assessment of Competence in Monitoring Student Achievement in the Classroom to $134 per teacher for a single observation using the Science Laboratory Assessment. In general, if the developmental work recommended is done, the less expensive assessments would be suitable for licensure decisions based on the teaching competencies measured by the assessment instrument, but produce limited diagnostic information to inform staff development choices.

Policy Issues

The ultimate goal of the California New Teacher Project is not the identification of better assessment instruments, but the improvement of teaching in schools. Thus, the major criterion by which the assessment approaches which have been pilot tested should be judged is their cost-effectiveness in improving the instruction in the California public schools. Obviously, these assessment instruments cannot bear this burden alone. If an additional credentialing requirement involving one or more of these tests is implemented, it will need to be coordinated with other reform efforts, such as the implementation of the subject-specific Curriculum Guides and Frameworks and the California Standards for Beginning Teachers as well as any statewide programs of new teacher support.

Evaluation of the cost-effectiveness of different assessment approaches in improving instruction, and the identification of ancillary reforms needed to insure this effectiveness, requires consideration of policy issues beyond the scope of this report, which focusses on the strengths and weaknesses of individual instruments and assessment approaches in measuring teaching competencies. The *Assessment Component of the California New Teacher Project: First Year Report* contained recommendations for policy decisions needed to guide the choice of assessment instruments as an additional credentialing requirement for new teachers. Our experience with the pilot tests described in this report leads us to propose the following additions and/or revisions to that list:

- **Assessment focus**. In the first year report, we identified this area as a major decision to be made in the design of an assessment system. In this report, we wish to augment that recommendation. We have pilot tested five assessments in the second year of the project in addition to the four pilot tested during the first year, and have yet to identify any one assessment approach which does not exhibit a weakness in at least one important area of teacher competence. Multiple assessment approaches are needed to assess a wide range of competencies or the state will need to identify one or two areas of major interest (e.g., content pedagogy, classroom management).

- **Appropriateness for Beginning Teachers**. When teacher performance across multiple assessments is examined, some common weaknesses appear in the following areas: sequencing instruction, choosing appropriate representations of content, the breadth and depth of content knowledge, and designing instruction for different types of students. Given that a literature on the development of teaching skills is just beginning to emerge, we do not know if these weaknesses are developmental weaknesses typical of beginning teachers or are results of weaknesses in the current curriculum for the preparation of teachers. How are new teachers to develop these skills? If teacher preparation programs need to change, what is the incentive, and how can the programs be assisted in their efforts to change, as well as monitored to ensure that change occurs? What would be done with teachers whose preparation occurred in another state? If it is expected that these skills develop on the job with additional experience, what policies and/or programs could facilitate their development? Are more complex assessments of new teachers worth implementing if these skills are ignored?

- **Coordination with professional development.** In a previous report (Estes et al., 1990), we noted that one decision to be made with respect to the design of an assessment system was the extent to which credentialing assessments should be coordinated with staff development activities. We can now better describe the impact of that decision on the choice of assessments. Based on our comparison of assessment approaches, we note that some approaches, such as Classroom Observations, would be well suited for providing information to guide staff development, while others, such as Structured Simulation Tasks, would not.

- **Teaching of diverse students.** A teaching credential licenses a teacher to teach in any classroom in the state. However, beginning teachers possess in-depth knowledge of only a limited range of students, chiefly those experienced in student teaching and the first year(s) of teaching. Assessments which tap a depth of knowledge of teaching probably need to focus on the students with whom the teacher is familiar. Caution needs to be taken with assessments which focus on teaching specific types of students, as these might differentially advantage teachers according to the degree of experience with each type of student, making fairness difficult to maintain. In addition, thought needs to be given to the identification and retention of teachers who excel in contexts where student achievement is typically low, even if they do not possess the breadth of knowledge of students which would enable them to teach effectively in other teaching contexts. However, licensing teachers to teach in specific contexts as well as in specific content areas would be a logistical nightmare. The degree to which the limited experience of beginning teachers is balanced against an interest in assessing a teacher's ability to teach in multiple teaching contexts is an important policy issue.

The policy decisions outlined in this report and the previous report will affect the design of any system for assessing new teachers, which will be contained in a report to the State Legislature in 1992. The present and future reports analyzing the pilot testing conducted in the three years of the Assessment Component of the California New Teacher Project provide information on the strengths and weaknesses of a number of different approaches to teacher assessment as well as specific instruments representing these approaches.

APPENDIX A:

STATISTICAL COMPARISON OF TEACHER PERFORMANCE ON THE SECONDARY LIFE/GENERAL SCIENCE TEACHER ASSESSMENT

```
                                                            17:14 Thursday, October 18, 1990  15

                            CTC PILOT TEST ANALYSES: RAND DATA
                         Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                       Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                        Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                           Analysis A: Descriptive Statistics
                 Part 4: Task Level Scores summed across raters by FORM

------------------------------------- Form=Form A -----------------------------------------------------
                         Applying Effective            Teacher as Curriculum
                       Instructional Techniques           Decision-Maker              Parent/Student Letter
  ---------------------------------------------------------------------------------------------------------
 |                        |T1SS: Task 1, total, Scorer |T2SS: Task 2, total, Scorer |T3SS: Task 3, total, Scorer |
 |                        |            Sum             |            Sum             |            Sum             |
 |                        |----------------------------+----------------------------+----------------------------|
 |                        | MEAN | STD  | N  |MIN |MAX | MEAN | STD  | N  |MIN |MAX | MEAN | STD  | N  |MIN |MAX |
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
A.1 Candidate's gender    |      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |------------------------|      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |- Male                  |  7.36|  1.75|  11|   4|  10| 72.91| 12.91|  11|  44|  90| 13.09|  6.95|  11|   2|  22|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |- Female                |  7.00|  2.19|  21|   0|  10| 77.10| 12.81|  21|  50| 100| 14.76|  5.79|  21|   2|  24|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |Candidate's race        |      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |------------------------|      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |- Non-white             |  7.25|  3.06|   8|   0|  10| 81.75|  5.28|   8|  72|  88| 17.25|  6.18|   8|   9|  24|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |- White                 |  7.08|  1.64|  24|   4|  10| 73.62| 13.96|  24|  44| 100| 13.17|  5.92|  24|   2|  23|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |Teacher's prep          |      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |------------------------|      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |- Missing               |  0.00|     .|   1|   0|   0| 72.00|     .|   1|  72|  72|  9.00|     .|   1|   9|   9|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |- Regular               |  7.46|  1.72|  24|   4|  10| 74.42| 13.79|  24|  44| 100| 13.79|  6.52|  24|   2|  24|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |- Intern                |  7.00|  1.00|   7|   6|   9| 80.43|  9.05|   7|  66|  89| 16.29|  4.72|   7|  11|  22|
 |------------------------+------+------+----+----+----+------+------+----+----+----+------+------+----+----+----|
 |Candidate's grade       |      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |------------------------|      |      |    |    |    |      |      |    |    |    |      |      |    |    |    |
 |- Middle                |  7.70|  1.25|  10|   6|  10| 79.00| 16.10|  10|  44| 100| 15.30|  6.86|  10|   6|  24|
  ---------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                              17:14 Thursday, October 18, 1990  16

                                        CTC PILOT TEST ANALYSES: RAND DATA
                                       Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                                    Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                                     Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                                        Analysis A: Descriptive Statistics
                                   Part 4: Task Level Scores summed across raters by FORM

--------------------------------------------------------- Form=Form A ---------------------------------------------------------

  -------------------------------------------------------------------------------------------------------------------------
 |                               | T1SS: Task 1, total, Scorer  | T2SS: Task 2, total, Scorer  | T3SS: Task 3, total, Scorer  |
 |                               |             Sum              |             Sum              |             Sum              |
 |                               |------------------------------+------------------------------+------------------------------|
 |                               | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX |
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |Candidate's grade              |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |- Junior high school           |   6.00|   2.93|   8|   0|  10|  74.87|   9.22|   8|  61|  90|  15.00|   5.88|   8|   8|  23|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |- Continuous, Regular, JHS/HS  |   7.36|   1.74|  14|   4|  10|  73.71|  12.33|  14|  50|  90|  12.93|   5.99|  14|   2|  23|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |Candidate's grade, collapsed   |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |- Middle, Junior high school   |   6.94|   2.26|  18|   0|  10|  77.17|  13.29|  18|  44| 100|  15.17|   6.26|  18|   6|  24|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |- Continuous, Regular, JHS/HS  |   7.36|   1.74|  14|   4|  10|  73.71|  12.33|  14|  50|  90|  12.93|   5.99|  14|   2|  23|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |Candidate's location           |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |- Inner City                   |   7.00|   1.50|   9|   4|   9|  82.33|  10.02|   9|  64| 100|  16.56|   5.96|   9|   6|  24|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |- Others                       |   7.17|   2.23|  23|   0|  10|  73.04|  12.99|  23|  44|  97|  13.26|   6.10|  23|   2|  23|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |Candidate's age                |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
 |- 29 or below                  |   7.72|   1.45|  18|   5|  10|  76.22|   9.85|  18|  60|  97|  13.83|   6.64|  18|   2|  23|
 |-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
 |- 30-34                        |   6.82|   1.78|  11|   4|  10|  76.64|  14.79|  11|  50| 100|  15.00|   5.22|  11|   6|  24|
  -------------------------------------------------------------------------------------------------------------------------

 (CONTINUED)
```

```
                                                                  17:14 Thursday, October 18, 1990  17

                            CTC PILOT TEST ANALYSES: RAND DATA
                         Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                      Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                       Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                           Analysis A: Descriptive Statistics
                 Part 4: Task Level Scores summed across raters by FORM

------------------------------------------------------- Form=Form A -------------------------------------------------------

  ------------------------------------------------------------------------------------------------------------------------
  |                              | T1SS: Task 1, total, Scorer | T2SS: Task 2, total, Scorer | T3SS: Task 3, total, Scorer |
  |                              |             Sum             |             Sum             |             Sum             |
  |                              |-----------------------------+-----------------------------+-----------------------------|
  |                              | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX |
  |------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
A.3|Candidate's age              |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
  |------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
  |- 35 or above                 |   4.67|   4.16|   3|   0|   8|  68.67|  23.18|   3|  44|  90|  13.33|   8.39|   3|   8|  23|
  |------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
  |ALL                           |   7.12|   2.03|  32|   0|  10|  75.66|  12.79|  32|  44| 100|  14.19|   6.15|  32|   2|  24|
  ------------------------------------------------------------------------------------------------------------------------
```

BEST COPY AVAILABLE

```
                                                                  17:14 Thursday, October 18, 1990  18

                              CTC PILOT TEST ANALYSES: RAND DATA
                           Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                         Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                          Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                              Analysis A: Descriptive Statistics
                      Part 4: Task Level Scores summed across raters by FORM

---------------------------------------------------- Form=Form B ---------------------------------------------------
                                      Applying Effective                                        Classroom and
 -------------------------------Instructional Techniques----------Lesson Planning---------------Facility Safety----
|                               |T1SS: Task 1, total, Scorer   |T2SS: Task 2, total, Scorer   |T3SS: Task 3, total, Scorer   |
|                               |             Sum              |             Sum              |             Sum              |
|                               +------------------------------+------------------------------+------------------------------|
|                               | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX |
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|Candidate's gender             |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|- Male                         |   8.86|   3.21|  14|   4|  14|   3.29|   1.44|  14|   0|   5|  12.46|   3.20|  13|   7|  16|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|- Female                       |   8.11|   2.45|  19|   4|  12|   4.05|   1.54|  19|   2|   7|  13.05|   3.32|  19|   6|  18|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|Candidate's race               |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|- Missing                      |   8.00|      .|   1|   8|   8|   3.00|      .|   1|   3|   3|  10.00|      .|   1|  10|  10|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|- Non-white                    |   8.00|   3.42|   8|   4|  14|   3.62|   1.60|   8|   2|   6|  11.38|   3.34|   8|   6|  16|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|- White                        |   8.58|   2.65|  24|   4|  14|   3.79|   1.56|  24|   0|   7|  13.43|   3.12|  23|   7|  18|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|Teacher's prep                 |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|-------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
|- Regular                      |   8.54|   2.72|  24|   4|  14|   3.71|   1.65|  24|   0|   7|  13.39|   3.16|  23|   7|  18|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|- Intern                       |   7.29|   2.75|   7|   4|  11|   3.71|   1.38|   7|   2|   5|  10.43|   2.82|   7|   6|  14|
|-------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
|- Other                        |  11.00|   2.83|   2|   9|  13|   4.00|   0.00|   2|   4|   4|  14.50|   2.12|   2|  13|  16|
 ----------------------------------------------------------------------------------------------------------------------
```

(CONTINUED)

```
                                                                      17:14 Thursday, October 18, 1990  19

                                  CTC PILOT TEST ANALYSES: RAND DATA
                                Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                              Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                               Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                                  Analysis A: Descriptive Statistics
                           Part 4: Task Level Scores summed across raters by FORM

------------------------------------------------------------ Form=Form B --------------------------------------------------------------

  --------------------------------------------------------------------------------------------------------------------------------
 |                                  |T1SS: Task 1, total, Scorer  |T2SS: Task 2, total, Scorer  |T3SS: Task 3, total, Scorer  |
 |                                  |            Sum              |            Sum              |            Sum              |
 |                                  |-----------------------------+-----------------------------+-----------------------------|
 |                                  | MEAN | STD  | N  |MIN |MAX  | MEAN | STD  | N  |MIN |MAX  | MEAN | STD  | N  |MIN |MAX   |
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |Candidate's grade                 |      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |----------------------------------|      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |- Middle                          |  7.86|  1.86|   7|   6|   10|  4.00|  1.91|   7|   2|    7| 12.14|  3.39|   7|   7|    16|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |- Junior high school              |  8.25|  3.33|  12|   4|   14|  3.67|  1.30|  12|   2|    6| 11.09|  3.21|  11|   6|    16|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |- Continuous, Regular, JHS/HS     |  8.86|  2.74|  14|   4|   14|  3.64|  1.60|  14|   0|    6| 14.50|  2.44|  14|   8|    18|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |Candidate's grade, collapsed      |      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |----------------------------------|      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |- Middle, Junior high school      |  8.11|  2.83|  19|   4|   14|  3.79|  1.51|  19|   2|    7| 11.50|  3.22|  18|   6|    16|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |- Continuous, Regular, JHS/HS     |  8.86|  2.74|  14|   4|   14|  3.64|  1.60|  14|   0|    6| 14.50|  2.44|  14|   8|    18|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |Candidate's location              |      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |----------------------------------|      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |- Inner City                      |  8.33|  3.24|   9|   4|   14|  3.22|  1.09|   9|   2|    5| 12.13|  3.48|   8|   7|    16|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |- Others                          |  8.46|  2.65|  24|   4|   14|  3.92|  1.64|  24|   0|    7| 13.04|  3.20|  24|   6|    18|
 |----------------------------------+------+------+----+----+-----+------+------+----+----+-----+------+------+----+----+------|
 |Candidate's age                   |      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |----------------------------------|      |      |    |    |     |      |      |    |    |     |      |      |    |    |      |
 |- 29 or below                     |  9.06|  2.43|  16|   4|   14|  3.62|  1.63|  16|   0|    6| 12.69|  2.91|  16|   6|    16|
  --------------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                                17:14 Thursday, October 18, 1990  20

                                            CTC PILOT TEST ANALYSES: RAND DATA
                                        Program: F:\DATA\SAS\PROGRAMS\RAND_A.SAS
                                     Input Data: F:\DATA\SAS\DATASETS\RAND_DAT.SSD
                                      Output File: F:\DATA\SAS\OUTPUT\RAND_A.OUT

                                            Analysis A: Descriptive Statistics
                                   Part 4: Task Level Scores summed across raters by FORM

----------------------------------------------------- Form=Form B ----------------------------------------------------

   -------------------------------------------------------------------------------------------------------------------
   |                              |T1SS: Task 1, total, Scorer |T2SS: Task 2, total, Scorer |T3SS: Task 3, total, Scorer |
   |                              |            Sum             |            Sum             |            Sum             |
   |                              |----------------------------+----------------------------+----------------------------|
   |                              | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX | MEAN  |  STD  | N  |MIN |MAX |
   |------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
A.6|Candidate's age               |       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
   |------------------------------|       |       |    |    |    |       |       |    |    |    |       |       |    |    |    |
   |- 30-34                       |   6.17|   2.04|   6|   4|   9|   3.83|   1.60|   6|   2|   6|  15.33|   2.07|   6|  12|  18|
   |------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
   |- 35 or above                 |   8.73|   3.13|  11|   4|  14|   3.82|   1.47|  11|   2|   7|  11.50|   3.66|  10|   7|  16|
   |------------------------------+-------+-------+----+----+----+-------+-------+----+----+----+-------+-------+----+----+----|
   |ALL                           |   8.42|   2.77|  33|   4|  14|   3.73|   1.53|  33|   0|   7|  12.81|   3.24|  32|   6|  18|
   -------------------------------------------------------------------------------------------------------------------
```

APPENDIX B:

SCIENCE LABORATORY ASSESSMENT: CONTENT AND FORMS

California New Teacher Project
Science Laboratory Assessment

Domains, Elements, & Indicators

RMC Research Corporation
Mountain View, California

A. Pedagogy

A1. Planning - The objective(s) for the activity involves the development or utilization of one or more of the scientific thinking processes (i.e., observing, communicating, comparing, ordering, categorizing, relating, inferring, and applying). The objective(s) and the activity are not focused solely on facts, but also on concepts and processes. The teacher knows what prerequisite skills and knowledge are required for the planned activity and the extent to which the students have these. The activity is an appropriate one for helping students to achieve the objective(s) and one that can be safely implemented with the students, given the facilities, equipment, and materials that are available.

A2. Sequence - The teacher organizes the steps or tasks of the laboratory activity in a logical or purposeful manner that allows students to achieve the lesson objective(s) and to complete the activity in an effective manner. Students do not exhibit confusion, or incorrect procedures or conclusions that might be due to inappropriate sequencing of the steps and tasks in the activity.

A3. Prelab - The teacher provides the students with a focus for the activity and a framework for learning. The opening is related to the teacher's objective(s). The teacher may: explain the purpose of the activity, help the students anticipate the activity, link it to the students' interests, point out the relevance of the activity to the students' own lives, tie the activity to prior learning in the class or to other subject areas, provide motivation for the students to learn from the lab activity, or review background materials.

A4. Directions - The teacher provides clear and comprehensive directions, orally and/or in writing as needed, to the students for doing the activity. The directions are at an appropriate level of complexity and difficulty for the students. The teacher communicates expectations for students' work on the activity.

A5. Explanation/Presentation - The teacher provides clear and accurate explanations, presentations, and reviews of concepts, principles, definitions, and processes, as needed,

BEST COPY AVAILABLE

or allows students to do these. The teacher may do a brief demonstration or modeling of the activity, as appropriate for the students and the lesson objectives. The explanations/presentations are at a suitable level of complexity and difficulty for the students, are visible to all students, and are related to the objective(s) of the activity.

A6. Monitoring/Adjusting - The teacher monitors student understanding and work during the activity; usually this involves walking around the room during the activity. Monitoring may consist of asking students specific questions about their understanding of the content, methods or equipment; observing students doing the activity; and listening to students discuss the activity with each other and with the teacher. During the observation period, the teacher adjusts the lesson or activity for individual students, small groups, or class-as-a-whole, as needed. Adjusting can take such forms as providing guidance, reviewing lesson content, presenting the information in a different manner, clarifying information, modeling a step, or changing the sequencing of the steps or tasks.

A7. Feedback - The teacher provides immediate, appropriate, and uncritical feedback to all students, individually or as a group, to promote attainment of the instructional objectives. Feedback (including feedback on wrong answers and errors) provides positive rewards, useful information, further motivation, or encouragement to students. The teacher provides objective feedback to students regardless of ability, ethnicity, gender, or other characteristics. The teacher's feedback indicates that the teacher can distinguish among a student's response that represents a nonstandard but productive insight, a response that indicates confusion, and one that indicates apathy.

A8. Questioning - The teacher asks questions that promote higher-order thinking processes and achievement of the objectives. Questions can be directed to the entire class, small groups, or individual students. The teacher involves as many students as possible, regardless of gender, ethnicity, language proficiency, or intellectual ability. The teacher asks questions at an appropriate cognitive level to encourage the development of skills in one or more of these processes:

- observing
- communicating
- comparing
- ordering
- categorizing
- relating
- inferring
- applying

The teacher provides appropriate wait time for students to respond. The teacher encourages the students to develop their own questions and answers, and builds on student responses and comments.

A9. Closure - When appropriate and in keeping with the objectives of the activity, the teacher provides a "wrap-up" or summary, and links the lesson objectives to past or future learning, or allows students to provide such closure. The teacher may review and summarize the purpose of the activity and what conclusions can be drawn from it, or allows students to provide such closure. Closure is often done at the end of a lab activity, but could also be done at the end of major segments completed earlier in the lab period or in a succeeding class period. The closure is directly related to the objective(s) of the lab activity.

B. Content

B1. Accurate - The teacher presents information that is accurate, and uses scientific content, methods and procedures that are generally accepted in the scientific community.

B2. Integrated - The teacher knows how the topic of the activity is connected to and interrelated with a major theme of science (e.g., energy, patterns of change, stability), other scientific topics, and topics in other subject areas. The teacher provides the means by which students can interrelate and connect the topic of the activity to: (1) past and future learnings on this topic, (2) other scientific topics, and/or (3) topics and academic skills in other subject areas.

B3. Related to Objectives - The teacher presents scientific information and uses methods that are related to the objective(s) of the laboratory activity.

C. Materials/Equipment

C1. Teacher Use - The teacher properly uses the equipment and handles the materials employed in the observed laboratory activity. Live organisms are maintained and handled in a humane and appropriate manner. Where applicable, the teacher is alert to student allergies, fears, and other problems related to the use of specimens or live organisms in the science lab activity.

C2. Safe Setup - The setup of equipment, furniture, and materials has no serious irregularities or dangerous conditions. The setting has, as needed, adequate ventilation, first aid supplies, safety equipment, corrosive-resistant counter tops, a fire extinguisher, running water, good lighting, etc. Materials and equipment are stored, labeled, and moved properly.

C3. Safe Practices - The teacher knows about the potential dangers involved in the planned science laboratory activity. The teacher informs students about, checks for understanding of, and enforces the proper use of equipment and handling of materials, as needed. The teacher tells students about safety procedures, potential dangers and actions to take, and proper cleanup and disposal procedures. Students are wearing safety gear (e.g., goggles, aprons, gloves) when needed. Cleanup and disposal are completed in a well-coordinated and safe manner. The teacher is alert to potential safety problems, knows what to do if a safety problem occurs, and takes corrective measures when necessary. There are no observed teacher violations of state and federal safety laws and regulations on the setup, use, and handling of materials and equipment.

C4. Availability - The teacher has provided a sufficient supply of materials and the necessary equipment so that all students can complete the activity and attain the lesson objective(s). The teacher prepares and modifies the equipment to be appropriate for the students (e.g., pre-mixes solutions, ties strings on weights). The teacher has all necessary materials and equipment for the lab activity available and ready to be used prior to the start of the class period. The procedures for distributing and the placement of equipment are suitable for the facilities or layout of the class setting. The teacher has provided students with easy and orderly access to the materials and equipment. Provisions have been made for physically disabled students, when present, so that they are able to participate in a meaningful way.

D. Management

D1. Grouping - The grouping of students is done in a manner that facilitates the completion of the activity, and the learning of the instructional objectives. The teacher has considered such factors as suitable size for the activity, the number and locations of work stations, the amount of equipment and working space available, the time needed to do each step, the roles assigned to various members of each group, the variable work rate of different students, and the grouping of students so they work well together.

D2. Other Personnel - If another person(s) (e.g., aide, peer tutor) is present and if that person is involved with the lab activity in an instructional or a managerial role, the teacher supervises that person's work as needed.

D3. Routines and Transitions - Classroom routines (e.g., taking attendance, distribution of materials, pencil sharpening) and transitions (e.g., from whole-class activity to small-group activity) occur smoothly and efficiently. There is continual progress toward students completing the activity and attaining the objective(s). The amount of non-instructional time is minimal.

D4. Student Engagement - The teacher structures the laboratory activity so that most of the students are engaged in a laboratory task most of the time. The teacher quickly attempts to reengage any student who is not on task or who deviates from the prescribed activity.

D5. Timing - The teacher allocates sufficient time for each step so that the students have an opportunity to complete the activity and attain the lesson objective(s). The teacher makes adjustments during the lab activity for students who complete it quickly, as well as for those who do not keep up with the other students.

D6. Student Behavior - The teacher encourages and reinforces appropriate student behavior. The teacher responds to student misbehavior quickly and positively. The teacher asserts control and maintains order so as to facilitate a productive lab activity.

D7. Lab Cleanup - Teacher devises, explains and implements lab cleanup procedures so that the lab setting is left neat and clean at the end of the activity.

E. Knowledge of Students

E1. Diversity - The teacher tailors instructional activities for a diverse classroom of students with different ethnic, cultural, language, and socioeconomic backgrounds and, when present, disabled students. The teacher does not compromise the rigor of the lesson and each student is challenged at an appropriate level.

E2. Student Characteristics - The teacher offers instruction and provides an activity that is appropriate for students' interests, cognitive and developmental levels, and prior knowledge. The teacher adjusts the information and activity for individual student differences when appropriate. The teacher is aware of student preconceptions and misconceptions that might interfere with the attainment of the objectives, and addresses them during the lab activity, as needed. The students seem to understand what is being taught and to be challenged by the activity and instruction.

F. Climate

F1. Interactions with Students - The teacher interacts with all students respectfully, positively, equitably, and in a culturally appropriate manner. The teacher is sensitive to students' preconceptions and values. The teacher avoids sarcasm and criticism. The teacher communicates high expectations for student learning and behavior, and provides all students with an opportunity to participate and learn.

F2. Interactions among Students - The teacher encourages and allows for productive and activity-related interactions and sharing among students. The students treat each other respectfully and politely.

F3. Attitudes - The teacher exhibits a positive attitude toward and enthusiasm for science. The teacher attempts to instill in students positive attitudes about learning and about science. The teacher demonstrates an attitude that the lab activity is a vital aspect of the students' learning, and that the individual student's results and observations are important.

F4. Inquiry - The teacher fosters an environment in which the processes of science are important, and an environment that promotes questioning, problem solving, discussion of error, and evaluation of competing ideas. The teacher does not place undue emphasis on students' obtaining "correct" or expected results in a laboratory activity. The teacher provides opportunities for discussion of anomalous results without embarrassing students. The teacher and students can criticize ideas without criticizing each other. The teacher encourages students to draw their own conclusions from observed data and to state them in their own words.

G. Communication

G1. Speaking - The teacher's oral communications (e.g., presentation, directions, feedback, informal conversations) are coherent and clear to all students. They have clarity of meaning and are given in a fluent manner and with a pleasant tone. Messages are not vague, ambiguous or incomplete. The teacher can be clearly heard by all students to whom the teacher is speaking (appropriate volume, enunciation, rate). Grammatical errors and mispronunciations, if any, are minimal and do not interfere with communication. The teacher does not use slang or vulgarities. The teacher uses acceptable conventions of spoken language for communicating with the students.

G2. Writing - The teacher's written communications (e.g., handouts, materials on chalkboard, overhead transparencies, posters made by the teacher, displays) are clear to students. They are not vague, ambiguous, or incomplete (unless required by the lesson objectives). Written materials can be clearly read by students (appropriate level of difficulty, legible, visible). Errors in grammar and spelling are minimal or nonexistent, and do not interfere with communication or set a bad example for students.

G3. Listening - The teacher listens to all students and reacts in an appropriate and supportive manner to their questions, answers, comments, failures to answer, errors, and needs.

G4. Strength of Presence - The teacher shows confidence with the science content, the methods and procedures, and the use of equipment and materials. When questions/situations arise that are outside of the teacher's background/experience, the teacher can frankly admit that and proceed to engage students in a cooperative effort to learn together. The teacher uses suitable body language and eye contact to gain and maintain control of the class and to hold the students' attention.

California New Teacher Project
Science Laboratory Assessment

Pre-Observation Questionnaire
(Part I of IV)

RMC Research Corporation
Mountain View, California

Instructions: *This form is to be completed by the teacher within 48 hours before being observed.*

Teacher______________________________

Date of Observation____-____-____

Principal______________________________

Date Questionnaire Completed___-___-___

School Name______________________________

Credential(s) Sought:

School Address______________________________

____

____ Multiple Subject (K-8)

____ Life Science

School Telephone (____)_______-_________

____ Physical Science

SCHEDULE	Location	Time
Pre-Observation Conference	____________________	___:___ to ___:___
Observation	____________________	___:___ to ___:___
Post-Observation Conference	____________________	___:___ to ___:___

Put this completed questionnaire and any relevant written materials (e.g., lesson plan, direction sheets for students, copy of students' data recording form) in an envelope, seal it, and mark the envelope as follows. Write your name in the space for teacher and the date of the scheduled observation on the line for date.

CNTP Science Lab Assessment

Teacher______________________________

Date____-____-____

Leave the sealed envelope in the school office. The observer for the CNTP pilot testing will pick it up early on the morning of the scheduled observation.

Section A. The Class Being Observed

These questions refer to the students who will be participating in the observed lab activity.

1. Name of Course____

2. Number of Students Enrolled______________ 3. Grade Level(s)__________

3a. If you have more than one grade level in this class, how many students are at each grade level? (e.g., 9th-14, 10th-15)

4. Are there any special needs students in this class (e.g., LEP, compensatory education, gifted, disabled) or any students who have behavior problems or are frequently disruptive? If yes, please list the number(s) of such students (e.g., 5 LEP, 3 gifted, 1 hearing impaired) and provide information that you think the observer should be aware of to understand what may be happening during the laboratory activity.

Numbers: ____

Other Information:

5. What is the general academic ability level of students in this class? (e.g., most at grade level, about 1/4 one grade lower; or, all college prep or honors)

6. Will any students be leaving or entering the room during the observation period?
Yes_____ No_____

6a. If yes, how many are there, and do they do this on a regular basis?

6b. How do they make up the time missed during the lab activity?

7. Will any other persons (e.g., bilingual aide, peer tutor) be present during the observation?
Yes_____ No_____ If yes, what will their roles be?

8. What administrative activities, not related to the laboratory activity, will occur during the observation period? (e.g., taking attendance)

9. Indicate the location of your lab activity. (Check one)

_____ Your regular classroom

_____ Your regular laboratory or classroom/lab combination

_____ Another classroom at your school

_____ Another laboratory or classroom/lab combination at your school

_____ Another location on the school site (e.g., yard)______________________________

_____ A location off the school site (e.g., beach, museum)___________________________

10. If the location is in your school building, do you share this location with other teachers? Yes_____ No_____

11. Is there anything else the observer should know about your classroom and/or the students?

Section B. The Laboratory Activity

12. Please complete the chart on the next page. List the major instructional objective(s), that is, what you want the students to be able to do as a result of this laboratory activity. For example: (1) Students will be able to weigh liquids, (2) Students will be able to calculate the density of liquids; or (1) Students will be able to focus a microscope, (2) Students will know how to draw a plant cell and an animal cell, showing the key structural parts, (3) Students will identify the structural parts associated with either plant cells or animal cells as well as those associated with both types of cells by comparing their two drawings.

In the appropriate space (or box) provided for each objective briefly describe each of the following:

- o the steps or tasks of the laboratory activity,
- o the student grouping planned,
- o the materials and equipment planned for that objective, and
- o the safety issues and precautions related to equipment and procedures.

California New Teacher Project
Science Laboratory Assessment

	Objective(s) (Student Outcomes)	Laboratory Activities (Steps, Tasks)	Student Groups (Size, Number of Stations, etc)	Materials and Equipment	Safety Issues Equipment and Procedures
1					
2					

Continue on the next page.

	Objective(s) (Student Outcomes)	**Laboratory Activities** (Steps, Tasks)	**Student Groups** (Size, Number of Stations, etc)	**Materials and Equipment**	**Safety Issues** Equipment and Procedures
3					
4					

Continue on back of this sheet if you need more space.

13. Source: What was your primary source(s) for this lab activity? Check those which apply.

____Textbook

____Professional journal or magazine

____Another teacher

____An outside science educational agency (e.g., museum, nature reserve)

____Developed solely by you

____Adapted by you from any of the above sources

____Other (please list)____

14. Theme: Which scientific theme(s) best pertains to your laboratory activity? See the descriptions on page 7. Check those which apply.

____Energy

____Environment

____Evolution

____Measurement

____Patterns of Change

____Scale and Structure

____Stability

____Systems and Interactions

____Other (please list)____

15. Content Area: Which content area(s) best reflects this laboratory activity? See the attached list on page 8. __________-__________

If other, specify area ____

16. Type: Which type of laboratory best categorizes this activity? (check one)

____Discovery/Inquiry

____Illustrative/Clarifying

____Exploratory

____Introductory

____Process Development

____Skills Development

____Other (please list)____

17. If there anything else the observer should know about the planned science laboratory activity, please write it below or on the back of this page.

Sign and date this form. Please see the instructions in the box on the first page.

Teacher's Signature__________________________Date Signed_____-_____-_____

List of Science Themes*

(for use with question 14)

1. Energy (capacity to do work or ability to make things move: the basis for reactions between chemical compounds: the ability of living systems to maintain their system. to grow. and to reproduce)

2. Environment (the surrounding circumstances and conditions: the impact of external or extrinsic conditions: appreciation of one's own environment; conservation: pollution)

3. Evolution (changes of natural entities and systems through time: the study of the patterns and processes that affect these changes)

4. Measurement (systems of measurement units: assessing dimensions, quantities, or capacities)

5. Patterns of Change (trends: cyclical patterns: irregular changes)

6. Scale and Structure (relationships of structures: hierarchical levels of structures and properties of each level: interplay of structure and function)

7. Stability (constancy: a balanced steady state: static and dynamic equilibrium)

8. Systems and Interactions (solar system: ecosystem: individual organisms: chemical and physical systems: input and output: feedback)

9. Other (specify theme)

* *Includes the six themes from the California Framework Draft (September 1989 Edition)*

List of Science Content Areas
(for use with question 15)

- I. Life Sciences
  - A. Cellular and Molecular Biology
  - B. Plants, Botany
  - C. Protists, Monerans, Fungi
  - D. Animals, Zoology
  - E. Human Beings
  - F. Ecosystems, Populations, Communities, Biogeochemical Cycles
  - G. Genetics
  - H. Evolution
  - I. Other (specify area)

- II. Earth Sciences
  - A. Basic Land and Water Forms
  - B. Structure of Rocks and Minerals
  - C. Structure of Solar System, Planetary Systems
  - D. Structure of Galaxies, the Universe
  - E. Movement of Materials (e.g., weathering, plates, tides)
  - F. Changes in Materials, Cycles (e.g., weather, rocks)
  - G. Evolution
  - H. Other (specify area)

- III. Physical Sciences--Chemistry
  - A. States of Matter
  - B. Models of Atomic, Molecular, Ionic Structures
  - C. Polarity and Implications for Properties of Molecules
  - D. Simple Nuclear Chemistry (e.g., radioactivity, fission, fusion)
  - E. Simple Models for Chemical Bonds (including implications for properties and geometries of molecules)
  - F. Factors that Govern Chemical Transformations (e.g., energy and spatial changes and effects, chemical equilibrium, electrolytic and voltaic cells, radioactive decay)
  - G. Chemical Reactions
  - H. Other (specify area)

- IV. Physical Sciences--Physics
  - A. Mechanics (e.g., motion, dynamics, gravity)
  - B. Conservation of Mass, Momentum and Energy
  - C. Heat
  - D. Electricity and Magnetism
  - E. Wave Motion (e.g., sound, light)
  - F. Atomic and Nuclear
  - G. Other (specify area)

- V. Safety and Manipulative Skills
  - A. Laboratory Safety
  - B. Manipulative Laboratory Skills

California New Teacher Project
Science Laboratory Assessment

Pre-Observation Conference Note-Taking Form

(Part II of IV)

Observer ______________________ Teacher ______________________ Date ______________

Start time ______________ End time ______________

Question#	Element Code(s)	Responses/Notes

California New Teacher Project
Science Laboratory Assessment

DOCUMENTATION SORTING RECORD

RMC Research Corporation
Mountain View, California

OBSERVER____________________ TEACHER____________________ DATE__________________

DOMAIN/ ELEMENT	EVIDENCE/NOTES/RESPONSES
A. PEDAGOGY:	
1. Planning	
2. Sequence	
3. Prelab	
4. Directions	
5. Explanation/ Presentation	

OBSERVER__________________ TEACHER __________________ DATE__________

DOMAIN/
ELEMENT EVIDENCE/NOTES/RESPONSES

6. Monitoring/
   Adjusting

7. Feedback

8. Questioning

9. Closure

B. CONTENT:

1. Accurate

2. Integrated

OBSERVER____________________ TEACHER____________________ DATE____________________

DOMAIN/ ELEMENT	EVIDENCE/NOTES/RESPONSES
3. Related to Objectives	
C. MATERIALS/EQUIPMENT:	
1. Teacher Use	
2. Safe Setup	
3. Safe Practices	
4. Availability	

OBSERVER________________ TEACHER________________ DATE________________

DOMAIN/
ELEMENT EVIDENCE/NOTES/RESPONSES

D. MANAGEMENT:
1. Grouping

2. Other
Personnel

3. Routines and
Transitions

4. Student
Engagement

5. Timing

580

OBSERVER____________________ TEACHER____________________ DATE____________________

DOMAIN/ ELEMENT	EVIDENCE/NOTES/RESPONSES
6. Student Behavior	
7. Lab Cleanup	
E. KNOWLEDGE OF STUDENTS:	
1. Diversity	
2. Student Characteristics	

OBSERVER____________________ TEACHER____________________ DATE________________

DOMAIN/
ELEMENT EVIDENCE/NOTES/RESPONSES

F. CLIMATE:

1. Interactions with Students

2. Interactions among Students

3. Attitudes

4. Inquiry

562

OBSERVER____________________ TEACHER____________________ DATE____________________

==

DOMAIN/
ELEMENT EVIDENCE/NOTES/RESPONSES

==

==

G. COMMUNICATION:

1. Speaking

2. Writing

3. Listening

4. Strength of
Presence

==

Observer:____
(sign)

Date:______/______/______

California New Teacher Project
Science Laboratory Assessment

SUMMARY REPORT FORM

TEACHER______________________________ DATE OF OBSERVATION______________________

JUDGMENTS	DOMAIN/ELEMENTS	REMARKS
☐	A. PEDAGOGY (Planning, Sequence, Prelab, Directions, Explanation/Presentati Monitoring/Adjusting, Feedback, Questioning, Closure)	
☐	B. CONTENT (Accurate, Integrated, Related to Objectives)	
☐	C. MATERIALS/EQUIPMENT (Teacher Use, Safe Setup, Safe Practices, Availability)	
☐	D. MANAGEMENT (Grouping, Other Personnel, Routines & Transitions, Stude Engagement, Timing, Student Behavior, Lab Cleanup)	

E. KNOWLEDGE OF STUDENTS (Diversity, Student Characteristics)

☐

F. CLIMATE (Interactions with Students, Interactions among Students, Attitudes, Inquiry)

☐

G. COMMUNICATION (Speaking, Writing, Listening, Strength of Presence)

☐

☐ OVERALL JUDGMENT

COMMENTS:

OBSERVER ______________________ DATE ____/____/____
printed

____
signed

Questions for the Pre-Observation Conference
(Part II of IV)

RMC Research Corporation
Mountain View, California

Below are the questions that will be asked during the Pre-Observation Conference.

1. I have reviewed your Questionnaire. Is there anything on it you need to change before we continue?

2. Why did you select this particular activity?

3. Did you design or modify the activity in order to make it appropriate for the students' background and interests, or to better enable you to accomplish your objectives? If yes, explain how and why.

4. Explain the scientific concepts and/or skills you are teaching in this lab activity.

5. What are some of the incorrect preconceptions that students may have that relate to this activity? *(pause)* How do you plan to address these during the lesson?

6. What prior instruction have you implemented related to the lab activity? *(pause)* What do students already know about this topic?

7. Have you provided previous instruction to ensure that students have the technical skills (e.g., students know how to use a voltmeter) requisite to the successful completion of this laboratory activity? If yes, was this provided recently? If not, what techniques have you employed to provide you evidence that students are ready to use the required processes and technical skills?

8. What instruction are you planning to do in the future related to the activity?

9. What is the relationship or contribution of this laboratory activity to the broad goals for the students' learning? *(pause)* Does it provide linkage from one concept to the next, or is it part of a continuing direction within one major concept? If yes, please explain.

10. What advanced thinking skills (e.g., comparing, estimating, inferring) will students be encouraged to use or required to apply in order to productively participate in this activity?

11. What factors did you consider in grouping students for this activity? *(pause)* Is this a departure from your normal grouping for this class?

12. What safety precautions will you take into consideration during this activity? *(pause)* What would you do if ________________________________? *(The observer should ask about a safety problem that might occur in this area of science; for example, a dangerous chemical spill, a heat burn, a deep cut on a student's hand.)*

13. Do you have a sufficient supply of materials and equipment for this activity? *(pause)* Are there any equipment problems or limitations that I should know about? *(pause)* If yes, how do you plan to cope with shortages or problems?

14. Are there any special procedures that must be followed in cleaning up after the activity?

15. Is there anything else you would like to tell me about your students and today's laboratory activity that we haven't covered in this meeting or that wasn't on the Questionnaire and that would help me better understand and assess the activity I observe?

California New Teacher Project
Science Laboratory Assessment

GUIDED NOTE-TAKING RECORD FOR THE OBSERVATION

Part III of IV

OBSERVER________________ TEACHER________________ DATE________ PAGE________
TIME________

DOMAIN/ ELEMENTS	EVIDENCE/NOTES (be specific)
A. PEDAGOGY 1. Planning 2. Sequence 3. Prelab 4. Directions 5. Explanation/Presentation 6. Monitoring/Adjusting 7. Feedback 8. Questioning 9. Closure	
B. CONTENT 1. Accurate 2. Integrated 3. Related to Objectives	
C. MATERIALS/ EQUIPMENT 1. Teacher Use 2. Safe Setup 3. Safe Practices 4. Availability	
D. MANAGEMENT 1. Grouping 2. Other Personnel 3. Routines & Transitions 4. Student Engagement 5. Timing 6. Student Behavior 7. Lab Cleanup	
E. KNOWLEDGE OF STUDENTS 1. Diversity 2. Student Characteristics	
F. CLIMATE 1. Interactions with Students 2. Interactions among Students 3. Attitudes 4. Inquiry	
G. COMMUNICATION 1. Speaking 2. Writing 3. Listening 4. Strength of Presence	

Questions for the Post-Observation Conference
(Part IV of IV)

RMC Research Corporation
Mountain View, California

Below are the questions to be asked during the Post-Observation Conference. The observer should add questions that will improve the understanding of what was observed and the assessment made for each Domain.

1. Did the lab activity go as you expected? *(pause)* If no, describe what happened that was unexpected. *(pause)* How are you going to deal with this problem?
2. Were the objectives attained by your students? *(pause)* What type of feedback will you provide to students now that the lab period is over?
3. How do you plan to assess the retention of these objectives?
4. Based on how your students did today, do you feel you need to do additional follow-up instruction related to this activity?
5. If you were to redo this activity, what changes would be desirable? *(Possible prompts: Any in the content? Any procedural changes? Any questions you might have asked students in order to redirect them? Any other changes?)*
6. *(If necessary, create your own question(s) to fill in missing information on the Domains and Elements. List your questions(s) on the Note-Taking Form.)*
7. *(If there are questions you have about areas not covered by the Domains and Elements, but which are related to the assessment process, ask these, too, and record the questions on the Note-Taking Form.)*
8. Is there anything else you would like to tell me about today's laboratory activity?

APPENDIX C:

AN EXAMPLE OF A SCORING SHEET FOR THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT

Question 2. Goals

What is the teacher's main goal for the students? Provide a rationale for why this is an important goal for a language arts activity.

(One point for the goal; one point for each supportive statement in the rationale; two points possible in the rationale section--for weaving two of the supportive statements into an answer; three points maximum)

A. *Main goal:*

__________ The primary goal of the teacher is to encourage her students to grapple with the challenge of making a difficult decision (decision making). The decision in this case (in the text and in the class activities) is one without a clear or single answer.

B. *Rationale:*

Note: Any item listed below could be used as a rationale for the goal.

The rationale for this goal as a language arts activity could include any or several of the following items:

__________ Goal is connected to the text (central character faces a similar decision)

__________ Students can benefit from examining/discussing the steps involved in decision making

__________ Students recognize there may be no one right or wrong choice, instead each alternative has good points and bad points

__________ Provides opportunity for meaningful, involved discussion of book

__________ Provides opportunity for discussion of ethical or cultural issues

__________ Discussion or class activities using this goal requires higher order thinking skills

__________ Class activities or discussions will incorporate listening, speaking, reading and writing (integrated approach)

__________ **Total for Question 2**

APPENDIX D:

STATISTICAL COMPARISON OF TEACHER PERFORMANCE ON THE LANGUAGE ARTS PEDAGOGICAL KNOWLEDGE ASSESSMENT

```
                                                                                  11:27 Monday, October 22, 1990  21

                                          CTC PILOT TEST ANALYSES: LAPKA DATA
                                       Program: F:\DATA\SAS\PROGRAMS\LAPKA_A.SAS
                              Input Data: F:\DATA\SAS\DATASETS\LAPKADAT.SSD, LAPKADEM.SSD
                                       Output File: F:\DATA\SAS\OUTPUT\LAPKA_A.OUT

                                         Analysis A: Descriptive Statistics
                      Part 4: Scenario and LAPKA Total Scores summed across raters (except for Scenario 3)

 ------------------------------------------------------------------------------------------------------------------------
|                              |           |Candidate's|   Candidate's     |Candidate's|Candidate's|  Candidate's    |     |
|                              |           |   race    |    education      |preparation|   grade   |   location      |     |
|                              |Candidate's|-----------|-------------------|-----------|-----------|-----------------|     |
|                              |  gender   |           |M=Da-|Non- |       |     |     |     |     |Subu-|     |     |     |
|                              |-----------|Non- |     | ta  |Cali-|Cali-|<= 1 |>= 2 |Grade|Grade|rban-|     |     |     |
|                              |     |Fema-|mino-|Mino-|Miss-|forn-|forn-|Cour-|Cour-|3 or |4 or |/Rur-|     |Inner|     |
|                              |Male | le  |rity |rity | ing | ia  | ia  | se  | ses |less |more | al  |Urban|City | ALL |
|------------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS1A_S: Total|MEAN          |    .|17.62|17.41|18.50|    .|17.50|17.65|17.33|17.83|17.82|17.40|18.43|16.67|18.20|17.62|
|for Scenario 1A|--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD           |    .| 3.54| 3.79| 2.38|    .| 1.29| 3.92| 4.97| 2.17| 4.73| 1.71| 2.70| 4.72| 1.92| 3.54|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N             |    0|   21|   17|    4|    0|    4|   17|    9|   12|   11|   10|    7|    9|    5|   21|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN           |    .| 5.00| 5.00|16.00|    .|16.00| 5.00| 5.00|14.00| 5.00|14.00|14.00| 5.00|15.00| 5.00|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX           |    .|22.00|22.00|21.00|    .|19.00|22.00|22.00|21.00|22.00|20.00|22.00|21.00|20.00|22.00|
|---------------+--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|PCTS1A_S: Pct. |MEAN          |    .|67.77|66.97|71.15|    .|67.31|67.87|66.67|68.59|68.53|66.92|70.88|64.10|70.00|67.77|
|for Scenario 1A|--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD           |    .|13.62|14.59| 9.16|    .| 4.97|15.08|19.13| 8.34|18.19| 6.59|10.38|18.14| 7.40|13.62|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N             |    0|   21|   17|    4|    0|    4|   17|    9|   12|   11|   10|    7|    9|    5|   21|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN           |    .|19.23|19.23|61.54|    .|61.54|19.23|19.23|53.85|19.23|53.85|53.85|19.23|57.69|19.23|
|               |--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX           |    .|84.62|84.62|80.77|    .|73.08|84.62|84.62|80.77|84.62|76.92|84.62|80.77|76.92|84.62|
|---------------+--------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS1B_S: Total|MEAN          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|for Scenario 1B|              | 7.25| 6.85| 6.92| 7.00| 7.00| 6.50| 7.08| 6.89| 7.00| 6.89| 7.00| 7.00| 6.83|    .| 6.94|
 ------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                 11:27 Monday, October 22, 1990   22

                          CTC PILOT TEST ANALYSES: LAPKA DATA
                        Program: F:\DATA\SAS\PROGRAMS\LAPKA_A.SAS
                 Input Data: F:\DATA\SAS\DATASETS\LAPKADAT.SSD, LAPKADEM.SSD
                      Output File: F:\DATA\SAS\OUTPUT\LAPKA_A.OUT

                          Analysis A: Descriptive Statistics
     Part 4: Scenario and LAPKA Total Scores summed across raters (except for Scenario 3)

-----------------------------------------------------------------------------------------------------------------------------
|                                    |           |           |   Candidate's   |Candidate's|Candidate's|   Candidate's   |     |
|                                    |           |Candidate's|    education    |preparation|   grade   |    location     |     |
|                                    |Candidate's|   race    |-----------------+-----------+-----------+-----------------|     |
|                                    |  gender   |-----------|M=Da-|Non- |     |     |     |     |     |Subu-|     |     |     |
|                                    |-----------|Non- |     | ta  |Cali-|Cali-|<= 1 |>= 2 |Grade|Grade|rban-|     |     |     |
|                                    |     |Fema-|mino-|Mino-|Miss-|forn-|forn-|Cour-|Cour-|3 or |4 or |/Rur-|     |Inner|     |
|                                    |Male | le  |rity |rity | ing | ia  | ia  | se  | ses |less |more | al  |Urban|City | ALL |
|------------------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS1B S: Total|STD                 | 0.96| 0.69| 0.79| 0.71|    .| 0.58| 0.79| 0.60| 0.93| 0.78| 0.76| 0.89| 0.41|    .| 0.75|
|for Scenario 1B|--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N                   |    4|   13|   12|    5|    1|    4|   12|    9|    8|    9|    8|   11|    6|    0|   17|
|               |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN                 | 6.00| 6.00| 6.00| 6.00| 7.00| 6.00| 6.00| 6.00| 6.00| 6.00| 6.00| 6.00| 6.00|    .| 6.00|
|               |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX                 | 8.00| 8.00| 8.00| 8.00| 7.00| 7.00| 8.00| 8.00| 8.00| 8.00| 8.00| 8.00| 7.00|    .| 8.00|
|---------------+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|PCTS1B S: Pct. |MEAN                |90.62|85.58|86.46|87.50|87.50|81.25|88.54|86.11|87.50|86.11|87.50|87.50|85.42|    .|86.76|
|for Scenario 1B|--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD                 |11.97| 8.61| 9.91| 8.84|    .| 7.22| 9.91| 7.51|11.57| 9.77| 9.45|11.18| 5.10|    .| 9.34|
|               |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N                   |    4|   13|   12|    5|    1|    4|   12|    9|    8|    9|    8|   11|    6|    0|   17|
|               |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN                 |75.00|75.00|75.00|75.00|87.50|75.00|75.00|75.00|75.00|75.00|75.00|75.00|75.00|    .|75.00|
|               |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX                 |100.0|100.0|100.0|100.0|87.50|87.50|100.0|100.0|100.0|100.0|100.0|100.0|87.50|    .|100.0|
|---------------+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS2 S: Total |MEAN                |53.25|46.03|47.70|43.90|64.00|47.13|46.10|49.06|45.04|44.32|48.95|47.55|44.67|49.80|46.75|
|for Scenario 2 |--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD                 | 7.72| 9.13| 9.68| 7.11|    .| 7.32| 9.30| 9.89| 8.42| 8.84| 9.11| 9.67| 9.52| 5.26| 9.17|
-----------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

BEST COPY AVAILABLE

```
                                                                                 11:27 Monday, October 22, 1990  23

                              CTC PILOT TEST ANALYSES: LAPKA DATA
                            Program: F:\DATA\SAS\PROGRAMS\LAPKA_A.SAS
                  Input Data: F:\DATA\SAS\DATASETS\LAPKADAT.SSD, LAPKADEM.SSD
                             Output File: F:\DATA\SAS\OUTPUT\LAPKA_A.OUT

                              Analysis A: Descriptive Statistics
       Part 4: Scenario and LAPKA Total Scores summed across raters (except for Scenario 1)

-----------------------------------------------------------------------------------------------------------------------------
|                                |             |             |    Candidate's     | Candidate's | Candidate's |    Candidate's     |      |
|                                |             | Candidate's |     education      | preparation |    grade    |      location      |      |
|                                | Candidate's |    race     |--------------------+-------------+-------------+--------------------|      |
|                                |   gender    |-------------|M=Da- | Non- |      |      |      |      |      |Subu- |      |      |      |
|                                |-------------| Non- |      |  ta  |Cali- |Cali- |<= 1  |>= 2  |Grade |Grade |rban- |      |      |      |
|                                |      |Fema- |mino- |Mino- |Miss- |forn- |forn- |Cour- |Cour- |3 or  |4 or  |/Rur- |      |Inner |      |
|                                |Male  |  le  |rity  |rity  | ing  |  ia  |  ia  |  se  | ses  |less  |more  |  al  |Urban |City  | ALL  |
|--------------------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|TOTS2 S: Total |N               |     4|    36|    30|    10|     1|     8|    31|    17|    23|    19|    21|    20|    15|     5|    40|
|for Scenario 2 |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |MIN             |46.00 |28.00 |28.00 |34.00 |64.00 |34.00 |28.00 |28.00 |28.00 |28.00 |28.00 |28.00 |28.00 |41.00 |28.00 |
|               |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |MAX             |64.00 |65.00 |65.00 |60.00 |64.00 |57.00 |65.00 |65.00 |63.00 |65.00 |64.00 |65.00 |63.00 |55.00 |65.00 |
|---------------+----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|PCTS2 S: Pct.  |MEAN            |64.94 |56.13 |58.17 |53.54 |78.05 |57.47 |56.22 |59.83 |54.93 |54.04 |59.70 |57.99 |54.47 |60.71 |57.01 |
|for Scenario 2 |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |STD             | 9.41 |11.13 |11.81 | 8.67 |     .| 8.92 |11.34 |12.06 |10.27 |10.79 |11.11 |11.79 |11.61 | 6.42 |11.19 |
|               |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |N               |     4|    36|    30|    10|     1|     8|    31|    17|    23|    19|    21|    20|    15|     5|    40|
|               |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |MIN             |56.10 |34.15 |34.15 |41.46 |78.05 |41.46 |34.15 |34.15 |34.15 |34.15 |34.15 |34.15 |34.15 |50.00 |34.15 |
|               |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |MAX             |78.05 |79.27 |79.27 |73.17 |78.05 |69.51 |79.27 |79.27 |76.83 |79.27 |78.05 |79.27 |76.83 |67.07 |79.27 |
|---------------+----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|TOTALS3: Total |MEAN            |26.25 |23.92 |24.06 |24.40 |30.00 |24.50 |23.87 |25.33 |23.22 |23.52 |24.80 |24.80 |23.12 |24.80 |24.15 |
|for Scenario 3,|----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|1 Rater        |STD             | 7.41 | 4.80 | 4.59 | 6.52 |     .| 3.66 | 5.34 | 4.73 | 5.18 | 5.88 | 4.02 | 5.61 | 4.79 | 3.42 | 5.04 |
|               |----------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|               |N               |     4|    37|    31|    10|     1|     8|    32|    18|    23|    21|    20|    20|    16|     5|    41|
-----------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                          11:27 Monday, October 22, 1990  24

                                  CTC PILOT TEST ANALYSES: LAPKA DATA
                                 Program: F:\DATA\SAS\PROGRAMS\LAPKA_A.SAS
                        Input Data: F:\DATA\SAS\DATASETS\LAPKADAT.SSD, LAPKADEM.SSD
                                Output File: F:\DATA\SAS\OUTPUT\LAPKA_A.OUT

                                   Analysis A: Descriptive Statistics
              Part 4: Scenario and LAPKA Total Scores summed across raters (except for Scenario 3)

 ----------------------------------------------------------------------------------------------------------
|                                   |           |           |   Candidate's     |Candidate's|Candidate's| Candidate's      |     |
|                                   |           |Candidate's|    education      |preparation|   grade   |  location        |     |
|                                   |Candidate's|   race    |-------------------+-----------+-----------+------------------|     |
|                                   |  gender   |-----------|M=Da-| Non-|     |     |     |     |     |Subu-|     |     |     |
|                                   |-----------| Non-|     | ta  |Cali-|Cali-|<= 1 |>= 2 |Grade|Grade|rban-|     |     |     |
|                                   |     |Fema-|mino-|Mino-|Miss-|forn-|forn-|Cour-|Cour-|3 or |4 or |/Rur-|     |Inner|     |
|                                   |Male | le  |rity |rity | ing | ia  | ia  | se  | ses |less |more | al  |Urban|City | ALL |
|-----------------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTALS3: Total   |MIN              |16.00|12.00|16.00|12.00|30.00|20.00|12.00|16.00|12.00|12.00|16.00|16.00|12.00|19.00|12.00|
|for Scenario 3,  |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|1 Rater          |MAX              |33.00|32.00|33.00|32.00|30.00|30.00|33.00|32.00|33.00|33.00|31.00|33.00|30.00|28.00|33.00|
|-----------------+-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|PCTS3: Pct. for  |MEAN             |75.00|68.34|68.76|69.71|85.71|70.00|68.21|72.38|66.34|67.21|70.86|70.86|66.07|70.86|68.99|
|Scenario 3, 1    |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|Rater            |STD              |21.17|13.73|13.11|18.62|    .|10.47|15.25|13.51|14.79|16.80|11.49|16.02|13.68| 9.77|14.39|
|                 |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |N                |    4|   37|   31|   10|    1|    8|   32|   18|   23|   21|   20|   20|   16|    5|   41|
|                 |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |MIN              |45.71|34.29|45.71|34.29|85.71|57.14|34.29|45.71|34.29|34.29|45.71|45.71|34.29|54.29|34.29|
|                 |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |MAX              |94.29|91.43|94.29|91.43|85.71|85.71|94.29|91.43|94.29|94.29|88.57|94.29|85.71|80.00|94.29|
|-----------------+-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS_S: Total    |MEAN             |86.75|84.47|85.81|81.44|101.0|83.62|84.44|86.94|82.74|81.06|88.39|84.50|82.29|94.25|84.72|
|LAPKA Score      |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |STD              |11.09|11.85|11.86|10.92|    .|13.38|11.11|12.30|10.96|10.90|11.47|11.58|12.21| 4.99|11.64|
|                 |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |N                |    4|   32|   27|    9|    1|    8|   27|   17|   19|   18|   18|   18|   14|    4|   36|
|                 |-----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                 |MIN              |74.00|63.00|63.00|66.00|101.0|63.00|66.00|66.00|63.00|68.00|63.00|63.00|66.00|90.00|63.00|
 ----------------------------------------------------------------------------------------------------------

(CONTINUED)
```

D.4

```
                         CTC PILOT TEST ANALYSES: LAPKA DATA
                    Program: F:\DATA\SAS\PROGRAMS\LAPKA_A.SAS
         Input Data: F:\DATA\SAS\DATASETS\LAPKADAT.SSD, LAPKADEM.SSD
                 Output File: F:\DATA\SAS\OUTPUT\LAPKA_A.OUT

                     Analysis A: Descriptive Statistics
 Part 4: Scenario and LAPKA Total Scores summed across raters (except for Scenario 3)

-------------------------------------------------------------------------------------------------------------------------
|                            |           |           |   Candidate's   |Candidate's|Candidate's|   Candidate's   |     |
|                            |           |Candidate's|    education    |preparation|   grade   |    location     |     |
|                            |Candidate's|   race    |-----------------+-----------+-----------+-----------------|     |
|                            |  gender   |-----------|M=Da-|Non- |     |     |     |     |     |Subu-|     |     |     |
|                            |-----------|Non- |     | ta  |Cali-|Cali-|<= 1 |>= 2 |Grade|Grade|rban-|     |     |     |
|                            |     |Fema-|mino-|Mino-|Miss-|forn-|forn-|Cour-|Cour-|3 or |4 or |/Rur-|     |Inner|     |
|                            |Male | le  |rity |rity | ing | ia  | ia  | se  | ses |less |more | al  |Urban|City | ALL |
|----------------+-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|TOTS_S: Total   |MAX        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|LAPKA Score     |           |101.0|105.0|105.0|97.00|101.0|101.0|105.0|105.0|105.0|105.0|105.0|105.0|105.0|101.0|105.0|
|----------------+-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|PCTS_S: Pct.    |MEAN       |69.40|62.15|63.45|61.46|80.80|62.15|62.53|64.97|61.15|60.46|65.45|64.30|60.38|65.91|62.96|
|LAPKA Score     |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |STD        | 8.87| 8.99| 9.32| 8.95|    .| 6.80| 9.35|10.67| 7.36|10.24| 7.35| 9.95| 8.97| 3.49| 9.14|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |N          |    4|   32|   27|    9|    1|    8|   27|   17|   19|   18|   18|   18|   14|    4|   36|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MIN        |59.20|47.55|47.55|49.65|80.80|50.40|47.55|47.55|47.55|47.55|50.40|47.55|47.55|62.94|47.55|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MAX        |80.80|84.00|84.00|77.60|80.80|70.63|84.00|84.00|73.43|84.00|80.80|84.00|77.60|70.63|84.00|
|----------------+-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|PCTAV_S: Ave.   |MEAN       |76.85|66.68|67.52|68.70|83.75|67.25|67.39|69.45|66.35|65.98|69.65|70.39|64.67|67.20|67.81|
|Pct. across     |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|Scenarios       |STD        | 6.37| 9.51| 9.72|10.18|    .| 4.54|10.54|11.49| 7.81|11.94| 6.64| 9.96| 9.99| 4.24| 9.70|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |N          |    4|   32|   27|    9|    1|    8|   27|   17|   19|   18|   18|   18|   14|    4|   36|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MIN        |68.90|43.48|43.48|53.80|83.75|61.53|43.48|43.48|53.80|43.48|59.76|55.20|43.48|63.59|43.48|
|                |-----------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MAX        |83.75|90.23|90.23|82.13|83.75|71.49|90.23|90.23|81.29|90.23|83.75|90.23|82.13|72.10|90.23|
-------------------------------------------------------------------------------------------------------------------------
```

APPENDIX E:

STATISTICAL COMPARISON OF TEACHER PERFORMANCE ON THE SECONDARY ENGLISH ASSESSMENT

E.1

```
                                                                      13:00 Sunday, October 14, 1990   1

                                  CTC PILOT TEST ANALYSES: SFSU DATA
                                Program: F:\DATA\SAS\PROGRAMS\SFSU_F.SAS
                             Input Data: F:\DATA\SAS\DATASETS\SFSU_DAT.SSD
                              Output File: F:\DATA\SAS\OUTPUT\SFSU_F.OUT

          Analysis F: Descriptive Statistics on Subtest and Major Test Ratings Summed Across Raters

----------------------------------------------------------------------------------------------------------------------
|                              |           |Candidate's|Candidate's preparation|Candidate's|Candidate's location       |     |
|                              |Candidate's|   race    |                       |   grade   |                           |     |
|                              |  gender   |-----------+-----------------------+-----------+---------------------------|     |
|                              |-----------| Non-|     |     | 0 or|     |     |     |Midd-|     |     |     |     |     |
|                              |     |Fema-|Mino-|Mino-|Miss-|  1  |  2  |  3  |High |le/J-|Miss-|Subu-|     |Inner|     |
|                              |Male | le  |rity |rity | ing |Cour-|Cour-|Cour-|Scho-| r.  | ing |rban |Urban|City | ALL |
|                              |     |     |     |     |     | ses | ses | ses | ol  |High |     |     |     |     |     |
|------------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SA1_RTS:        |MEAN         | 5.33| 6.08| 5.94| 5.33| 5.00| 5.29| 6.14| 6.50| 5.92| 5.71| 6.00| 6.50| 5.40| 5.57| 5.84|
|Overall         |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|response        |STD          | 0.82| 1.12| 1.12| 0.58|    .| 1.11| 0.69| 1.29| 0.79| 1.50|    .| 1.05| 0.89| 1.13| 1.07|
|strategies      |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |N            |    6|   13|   16|    3|    1|    7|    7|    4|   12|    7|    1|    6|    5|    7|   19|
|                |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MIN          |    4|    4|    4|    5|    5|    4|    5|    5|    5|    4|    6|    5|    4|    4|    4|
|                |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MAX          |    6|    8|    8|    6|    5|    7|    7|    8|    7|    8|    6|    8|    6|    7|    8|
|----------------+-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SA2_RTS:        |MEAN         | 4.75| 6.50| 6.10| 5.00|    .| 5.40| 6.00| 6.67| 6.12| 5.50| 6.00| 6.80| 5.50| 5.00| 5.92|
|Overall         |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|analysis of wrt |STD          | 0.96| 0.76| 1.10| 1.41|    .| 1.34| 0.82| 1.15| 1.25| 1.00|    .| 0.84| 0.71| 1.15| 1.16|
|and txt         |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |N            |    4|    8|   10|    2|    0|    5|    4|    3|    8|    4|    1|    5|    2|    4|   12|
|                |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MIN          |    4|    6|    4|    4|    .|    4|    5|    6|    4|    4|    6|    6|    5|    4|    4|
|                |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |MAX          |    6|    8|    8|    6|    .|    7|    7|    8|    8|    6|    6|    8|    6|    6|    8|
|----------------+-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|MA RTS: Overall |MEAN         | 5.17| 5.77| 5.69| 5.00| 5.00| 5.14| 5.86| 6.00| 5.83| 5.14| 6.00| 6.00| 5.20| 5.43| 5.58|
|rating, Form A  |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|                |STD          | 0.75| 1.01| 0.95| 1.00|    .| 0.90| 0.38| 1.63| 0.83| 1.07|    .| 1.26| 0.84| 0.79| 0.96|
----------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                              13:00 Sunday, October 14, 1990   2

                              CTC PILOT TEST ANALYSES: SFSU DATA
                           Program: F:\DATA\SAS\PROGRAMS\SFSU_F.SAS
                         Input Data: F:\DATA\SAS\DATASETS\SFSU_DAT.SSD
                          Output File: F:\DATA\SAS\OUTPUT\SFSU_F.OUT

        Analysis F: Descriptive Statistics on Subtest and Major Test Ratings Summed Across Raters

-------------------------------------------------------------------------------------------------------------------------
|                             |           |           |                       |Candidate's|                       |     |
|                             |           |Candidate's|Candidate's preparation|   grade   | Candidate's location  |     |
|                             |Candidate's|   race    |-----------------------+-----------+-----------------------|     |
|                             |  gender   |-----------|     |0 or |     |     |     |Midd-|     |     |     |     |     |
|                             |-----------|Non- |     |     |  1  |  2  |  3  |High |le/J-|     |     |     |     |     |
|                             |     |Fema-|Mino-|Mino-|Miss-|Cour-|Cour-|Cour-|Scho-| r.  |Miss-|Subu-|     |Inner|     |
|                             |Male | le  |rity |rity | ing | ses | ses | ses | ol  |High | ing |rban |Urban|City | ALL |
|-----------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|MA_RTS: Overall|N            |    6|   13|   16|    3|    1|    7|    7|    4|   12|    7|    1|    6|    5|    7|   19|
|rating, Form A |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN          |    4|    4|    4|    4|    5|    4|    5|    4|    5|    4|    6|    4|    4|    4|    4|
|               |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX          |    6|    8|    8|    6|    5|    6|    6|    8|    8|    6|    6|    8|    6|    6|    8|
|---------------+-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SB1_RTS:       |MEAN         | 6.67| 6.92| 6.69| 7.67| 6.00| 6.57| 7.14| 7.00| 7.33| 6.00| 8.00| 7.33| 6.40| 6.57| 6.84|
|Overall        |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|interpretive   |STD          | 1.03| 1.38| 1.30| 0.58|    .| 1.51| 1.07| 1.41| 0.98| 1.29|    .| 1.21| 1.67| 0.98| 1.26|
|process        |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N            |    6|   13|   16|    3|    1|    7|    7|    4|   12|    7|    1|    6|    5|    7|   19|
|               |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN          |    6|    4|    4|    7|    6|    4|    6|    5|    6|    4|    8|    5|    4|    6|    4|
|               |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX          |    8|    8|    8|    8|    6|    8|    8|    8|    8|    8|    8|    8|    8|    8|    8|
|---------------+-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SB2_RTS:       |MEAN         | 6.50| 6.85| 6.69| 7.00| 6.00| 6.86| 6.71| 6.75| 6.92| 6.43| 7.00| 7.33| 6.60| 6.29| 6.74|
|Overall group  |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|process        |STD          | 0.84| 0.99| 0.95| 1.00|    .| 1.07| 1.11| 0.50| 1.00| 0.79|    .| 0.82| 0.89| 0.95| 0.93|
|               |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N            |    6|   13|   16|    3|    1|    7|    7|    4|   12|    7|    1|    6|    5|    7|   19|
|               |-------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN          |    6|    5|    5|    6|    6|    6|    5|    6|    5|    6|    7|    6|    6|    5|    5|
-------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                             13:00 Sunday, October 14, 1990   1

                                          CTC PILOT TEST ANALYSES: SFSU DATA
                                        Program: F:\DATA\SAS\PROGRAMS\SFSU_F.SAS
                                      Input Data: F:\DATA\SAS\DATASETS\SFSU_DAT.SSD
                                       Output File: F:\DATA\SAS\OUTPUT\SFSU_F.OUT

                   Analysis F: Descriptive Statistics on Subtest and Major Test Ratings Summed Across Raters

-----------------------------------------------------------------------------------------------------------------------
|                                |           |           |                       |Candidate's|                       |     |
|                                |           |Candidate's|Candidate's preparation|   grade   | Candidate's location  |     |
|                                |Candidate's|   race    |-----------------------+-----------+-----------------------|     |
|                                |  gender   |-----------|     |0 or |     |     |     |Midd-|     |     |     |     |     |
|                                |-----------| Non-|     |     |  1  |  2  |  3  |High |le/J-|     |     |     |     |     |
|                                |     |Fema-|Mino-|Mino-|Miss-|Cour-|Cour-|Cour-|Scho-| r.  |Miss-|Subu-|     |Inner|     |
|                                |Male | le  |rity |rity | ing | ses | ses | ses | ol  |High | ing |rban |Urban|City | ALL |
|--------------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SB2_RTS:       |MAX             |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|Overall group  |                |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|process        |                |    8|    8|    8|    8|    6|    8|    8|    7|    8|    8|    7|    8|    8|    8|    8|
|---------------+----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|MB_RTS: Overall|MEAN            | 6.67| 7.08| 6.80| 7.67| 7.00| 6.86| 6.83| 7.25| 7.27| 6.43| 8.00| 7.40| 6.80| 6.57| 6.94|
|rating, Form B |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD             | 1.03| 1.08| 1.08| 0.58|    .| 1.07| 1.33| 0.96| 1.10| 0.79|    .| 0.89| 1.10| 1.13| 1.06|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N               |    6|   12|   15|    3|    1|    7|    6|    4|   11|    7|    1|    5|    5|    7|   18|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN             |    6|    5|    5|    7|    7|    6|    5|    6|    5|    6|    8|    6|    6|    5|    5|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX             |    8|    8|    8|    8|    7|    8|    8|    8|    8|    8|    8|    8|    8|    8|    8|
|---------------+----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SC1_RTS:       |MEAN            | 6.17| 6.38| 6.25| 6.67| 4.00| 5.71| 6.71| 7.25| 6.67| 5.71| 7.00| 6.83| 6.20| 5.86| 6.32|
|Overall content|----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |STD             | 1.47| 1.56| 1.61| 0.58|    .| 1.38| 1.50| 0.96| 1.30| 1.70|    .| 1.33| 1.30| 1.86| 1.49|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |N               |    6|   13|   16|    3|    1|    7|    7|    4|   12|    7|    1|    6|    5|    7|   19|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MIN             |    4|    4|    4|    6|    4|    4|    4|    6|    4|    4|    7|    5|    5|    4|    4|
|               |----------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|               |MAX             |    8|    8|    8|    7|    4|    8|    8|    8|    8|    8|    7|    8|    8|    8|    8|
-----------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

CTC PILOT TEST ANALYSES: SFSU DATA
Program: F:\DATA\SAS\PROGRAMS\SFSU_F.SAS
Input Data: F:\DATA\SAS\DATASETS\SFSU_DAT.SSD
Output File: F:\DATA\SAS\OUTPUT\SFSU_F.OUT

Analysis F: Descriptive Statistics on Subtest and Major Test Ratings Summed Across Raters

		Candidate's gender		Candidate's race		Candidate's preparation				Candidate's grade		Candidate's location				
		Male	Female	Non-Minority	Minority	Missing	0 or 1 Courses	2 Courses	3 Courses	High School	Middle/Jr. High	Missing	Suburban	Urban	Inner City	ALL
SC2_RTS: Overall plan	MEAN	6.00	6.08	6.13	5.67	4.00	5.86	6.17	6.75	6.27	5.71	6.00	6.83	5.50	5.71	6.06
	STD	0.71	1.50	1.41	0.58	.	1.21	1.17	1.50	1.19	1.50	.	1.33	1.29	1.25	1.30
	N	5	13	15	3	1	7	6	4	11	7	1	6	4	7	18
	MIN	5	4	4	5	4	4	4	5	4	4	6	5	4	4	4
	MAX	7	8	8	6	4	8	7	8	8	8	6	8	7	7	8
SC3_RTS: Overall delivery	MEAN	6.50	6.77	6.75	6.33	6.00	6.86	6.29	7.25	6.92	6.29	8.00	7.17	6.80	6.00	6.68
	STD	0.84	1.09	0.93	1.53	.	0.90	0.76	1.50	0.90	1.11	.	1.17	0.84	0.58	1.00
	N	6	13	16	3	1	7	7	4	12	7	1	6	5	7	19
	MIN	6	5	5	5	6	6	5	5	6	5	8	5	6	5	5
	MAX	8	8	8	8	6	8	7	8	8	8	8	8	8	7	8
MC RTS: Overall rating, Form C	MEAN	6.17	6.08	6.19	5.67	4.00	5.71	6.43	6.75	6.33	5.71	6.00	6.83	5.80	5.71	6.11
	STD	1.17	1.50	1.47	0.58	.	1.25	1.27	1.50	1.15	1.70	.	1.33	1.30	1.50	1.37

(CONTINUED)

E.5

```
                                                          13:00 Sunday, October 14, 1990   5

                         CTC PILOT TEST ANALYSES: SFSU DATA
                    Program: F:\DATA\SAS\PROGRAMS\SFSU_F.SAS
                 Input Data: F:\DATA\SAS\DATASETS\SFSU_DAT.SSD
                Output File: F:\DATA\SAS\OUTPUT\SFSU_F.OUT

    Analysis F: Descriptive Statistics on Subtest and Major Test Ratings Summed Across Raters

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|                                        |                   |                   |                                       |Candidate's        |                                       |         |
|                                        |                   |Candidate's        |Candidate's preparation                |       grade       |Candidate's location                   |         |
|                                        |Candidate's        |      race         |---------------------------------------|-------------------|---------------------------------------|         |
|                                        |      gender       |-------------------|         |0 or     |         |         |         |Midd-    |         |         |         |         |         |
|                                        |-------------------|Non-     |         |         |1        |2        |3        |High     |le/J-    |         |         |         |         |         |
|                                        |         |Fema-    |Mino-    |Mino-    |Miss-    |Cour-    |Cour-    |Cour-    |Scho-    |r.       |Miss-    |Subu-    |         |Inner    |         |
|                                        |Male     |le       |rity     |rity     |ing      |ses      |ses      |ses      |ol       |High     |ing      |rban     |Urban    |City     |ALL      |
|----------------------------------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
|MC RTS: Overall|N                       |        6|       13|       16|        3|        1|        7|        7|        4|       12|        7|        1|        6|        5|        7|       19|
|rating, Form C |------------------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
|               |MIN                     |        5|        4|        4|        5|        4|        4|        4|        5|        4|        4|        6|        5|        4|        4|        4|
|               |------------------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
|               |MAX                     |        8|        8|        8|        6|        4|        8|        8|        8|        8|        8|        6|        8|        7|        8|        8|
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

APPENDIX F:

STATISTICAL COMPARISON OF TEACHER PERFORMANCE ON THE SECONDARY ENGLISH ASSESSMENT: PORTFOLIO ACTIVITIES

```
                                                                        15:00 Monday, April 29, 1991   1

                                  CTC PILOT TEST ANALYSES: SFSU DATA
                              Program: D:\DATA\SAS\PROGRAMS\SFSU2E.SAS
                           Input Data: D:\DATA\SAS\DATASETS\SFSU2DAT.SSD
                            Output File: D:\DATA\SAS\OUTPUT\SFSU2E.OUT

               Analysis E: Descriptive Statistics on Subtest Ratings Summed Across Raters

--------------------------------------------------------------------------------------------------------------------------
|                           |           |           |Candidate's|                             |  Candidate's teacher prep   |
|                           |           |Candidate's|   grade   |    Candidate's location     |           school            |
|                           |Candidate's|   race    |-----------+-----------------------------+-----------------------------|
|                           |  gender   |-----------|     |Midd-|     |     |     |     |     |     |     |     |     |Priv-|
|                           |-----------| Non-|     |High |le/J-|     |     |     |     |     |     |     |     |Priv-|ate, |
|                           |     |Fema-|Mino-|Mino-|Scho-| r.  |Miss-|Subu-|     |Inner|     |Miss-|     |     |ate, | Not |
|                           |Male | le  |rity |rity | ol  |High | ing |rban |Urban|City |Rural| ing | CSU | UC  |in CA| CA  |
|--------------+------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SD1_RTS:      |MEAN        | 5.33| 4.80| 5.07| 4.50| 5.22| 4.71| 2.00| 4.40| 5.33| 6.00| 4.00| 2.00| 5.37| 6.33| 2.00| 4.67|
|Overall       |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|planning      |STD         | 1.97| 2.10| 2.06| 2.12| 1.86| 2.29|    .| 2.19| 2.08| 1.67|    .|    .| 1.92| 0.58|    .| 2.08|
|abilities     |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |N           |    6|   10|   14|    2|    9|    7|    1|    5|    3|    6|    1|    1|    8|    3|    1|    3|
|              |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |MIN         |    2|    2|    2|    3|    2|    2|    2|    2|    3|    4|    4|    2|    2|    6|    2|    3|
|              |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |MAX         |    7|    8|    8|    6|    8|    7|    2|    6|    7|    8|    4|    2|    8|    7|    2|    7|
|--------------+------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|SD2_RTS:      |MEAN        | 5.17| 4.60| 4.86| 4.50| 5.44| 4.00| 2.00| 5.20| 5.67| 4.33| 6.00| 2.00| 5.62| 6.67| 2.00| 2.67|
|Overall unit  |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|design skills |STD         | 2.23| 2.01| 2.11| 2.12| 1.94| 2.00|    .| 2.39| 2.52| 1.63|    .|    .| 1.41| 1.15|    .| 0.58|
|              |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |N           |    6|   10|   14|    2|    9|    7|    1|    5|    3|    6|    1|    1|    8|    3|    1|    3|
|              |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |MIN         |    2|    2|    2|    3|    2|    2|    2|    2|    3|    2|    6|    2|    4|    6|    2|    2|
|              |------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|              |MAX         |    8|    8|    8|    6|    8|    7|    2|    8|    8|    6|    6|    2|    8|    8|    2|    3|
--------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                          15:00 Monday, April 29, 1991   2

                                         CTC PILOT TEST ANALYSES: SFSU DATA
                                      Program: D:\DATA\SAS\PROGRAMS\SFSU2E.SAS
                                   Input Data: D:\DATA\SAS\DATASETS\SFSU2DAT.SSD
                                    Output File: D:\DATA\SAS\OUTPUT\SFSU2E.OUT

                    Analysis E: Descriptive Statistics on Subtest Ratings Summed Across Raters

-----------------------------------------------------------------------------------------------------------------------------------------------------------
|                                       |             | Candidate's | Candidate's |       Candidate's location       |     Candidate's teacher prep     |
|                                       |             |    race     |    grade    |                                  |              school              |
|                                       | Candidate's |-------------|-------------+----------------------------------+----------------------------------|
|                                       |   gender    |      |      |      |Midd- |      |      |      |      |      |      |      |      |      |Priv- |
|                                       |-------------| Non- |      | High |le/J- |      |      |      |      |      |      |      |      |Priv- |ate,  |
|                                       |      |Fema- |Mino- |Mino- |Scho- |  r.  |Miss- |Subu- |      |Inner |      |Miss- |      |      |ate,  | Not  |
|                                       |Male  | le   |rity  |rity  |  ol  |High  | ing  |rban  |Urban |City  |Rural | ing  | CSU  |  UC  |in CA | CA   |
|---------------------------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|SD3_RTS:            |MEAN              |  5.17|  5.30|  5.36|  4.50|  5.78|  4.57|  4.00|  5.40|  5.67|  5.17|  5.00|  4.00|  6.12|  6.67|  2.00|  3.00|
|Overall             |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|portfolio           |STD               |  1.47|  2.21|  1.95|  2.12|  1.64|  2.15|     .|  2.41|  2.31|  1.94|     .|     .|  1.46|  0.58|     .|  0.00|
|presentation        |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |N                 |     6|    10|    14|     2|     9|     7|     1|     5|     3|     6|     1|     1|     8|     3|     1|     3|
|                    |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |MIN               |     3|     2|     2|     3|     3|     2|     4|     2|     3|     3|     5|     4|     4|     6|     2|     3|
|                    |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |MAX               |     7|     8|     8|     6|     8|     8|     4|     8|     7|     8|     5|     4|     8|     7|     2|     3|
|--------------------+------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|SD4_RTS:            |MEAN              |  4.83|  5.00|  4.92|  5.00|  5.67|  3.83|  2.00|  5.00|  5.67|  5.00|  5.00|  2.00|  5.71|  6.67|  2.00|  3.33|
|Overall             |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|pedagogical         |STD               |  1.60|  2.00|  1.89|  1.41|  1.58|  1.60|     .|  2.16|  1.53|  1.79|     .|     .|  0.76|  0.58|     .|  1.15|
|abilities           |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |N                 |     6|     9|    13|     2|     9|     6|     1|     4|     3|     6|     1|     1|     7|     3|     1|     3|
|                    |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |MIN               |     2|     2|     2|     4|     2|     2|     2|     2|     4|     2|     5|     2|     5|     6|     2|     2|
|                    |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|                    |MAX               |     6|     7|     7|     6|     7|     6|     2|     7|     7|     7|     5|     2|     7|     7|     2|     4|
|--------------------+------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|SD5_RTS:            |MEAN              |  4.25|  4.70|  4.69|  3.00|  5.14|  4.00|  2.00|  5.40|  4.67|  4.20|     .|  2.00|  5.57|  6.00|  2.00|  3.00|
|Overall subject     |------------------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------|
|ped. abilities      |STD               |  1.71|  1.83|  1.75|     .|  1.68|  1.73|     .|  1.95|  1.53|  1.48|     .|     .|  0.79|  1.41|     .|  1.00|
-----------------------------------------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

```
                                                                          15:00 Monday, April 29, 1991   3

                                  CTC PILOT TEST ANALYSES: SFSU DATA
                                 Program: D:\DATA\SAS\PROGRAMS\SFSU2E.SAS
                               Input Data: D:\DATA\SAS\DATASETS\SFSU2DAT.SSD
                                Output File: D:\DATA\SAS\OUTPUT\SFSU2E.OUT

                     Analysis E: Descriptive Statistics on Subtest Ratings Summed Across Raters

---------------------------------------------------------------------------------------------------------------------------------------------
|                                |               |Candidate's    |Candidate's    |                                       |  Candidate's teacher prep             |
|                                |Candidate's    |     race      |     grade     |          Candidate's location         |            school                     |
|                                |    gender     |---------------+---------------|---------------------------------------+---------------------------------------|
|                                |---------------|Non-   |       |High   |Midd-  |       |       |       |       |       |       |       |       |       |Priv-  |
|                                |       |Fema-  |Mino-  |Mino-  |Scho-  |le/J-  |Miss-  |Subu-  |       |Inner  |       |Miss-  |       |       |Priv-  |ate,   |
|                                |Male   |le     |rity   |rity   |ol     |r.     |ing    |rban   |Urban  |City   |Rural  |ing    |CSU    |UC     |ate,   |Not    |
|                                |       |       |       |       |       |High   |       |       |       |       |       |       |       |       |in CA  |CA     |
|--------------------------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|SD5_RTS:        |N              |      4|     10|     13|      1|      7|      7|      1|      5|      3|      5|      0|      1|      7|      2|      1|      3|
|Overall subject |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|ped. abilities  |MIN            |      2|      2|      2|      3|      2|      2|      2|      2|      3|      2|      .|      2|      4|      5|      2|      2|
|                |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|                |MAX            |      6|      7|      7|      3|      7|      6|      2|      7|      6|      6|      .|      2|      6|      7|      2|      4|
|----------------+---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|SD6_RTS:        |MEAN           |   5.17|   5.20|   5.21|   5.00|   5.67|   4.57|   4.00|   5.80|   5.33|   5.17|   3.00|   4.00|   6.12|   6.67|   2.00|   2.67|
|Overall         |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|reflective      |STD            |   2.04|   2.10|   2.01|   2.83|   1.94|   2.07|      .|   2.17|   2.08|   2.23|      .|      .|   1.46|   0.58|      .|   0.58|
|ability         |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|                |N              |      6|     10|     14|      2|      9|      7|      1|      5|      3|      6|      1|      1|      8|      3|      1|      3|
|                |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|                |MIN            |      3|      2|      2|      3|      2|      2|      4|      2|      3|      2|      3|      4|      3|      6|      2|      2|
|                |---------------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------|
|                |MAX            |      7|      7|      7|      7|      7|      7|      4|      7|      7|      7|      3|      4|      7|      7|      2|      3|
---------------------------------------------------------------------------------------------------------------------------------------------

(CONTINUED)
```

APPENDIX G:

STATISTICAL COMPARISON OF GROUP PERFORMANCE ON THE ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT

TABLE

STATISTICAL COMPARISON OF GROUP PERFORMANCE ON THE ASSESSMENT OF COMPETENCE IN MONITORING STUDENT ACHIEVEMENT

	Pretest					
	Form A			Form B		
Group of Teachers	Mean	SD	N	Mean	SD	N
Males	39.8	12.7	5	42.6	11.7	9
Females	41.4	14.5	21	45.7	10.5	15
K-3 Teachers	46.1	14.4	12	46.3	8.8	12
4-6 Teachers	35.2	8.4	5	42.9	14.7	9
Inner-City Teachers	37.4	15.6	5	38.0	13.9	4
Non Inner-City Teachers	43.5	15.3	15	46.5	10.7	17
White Teachers	43.6	13.8	16	46.9	10.8	17
Minority Teachers	23.0	11.3	2	36.3	11.7	4

	Posttest					
	Form A			Form B		
Group of Teachers	Mean	SD	N	Mean	SD	N
Males	44.1	19.0	8	48.6	3.8	5
Females	43.9	13.8	14	47.7	12.7	19
K-3 Teachers	43.9	11.8	12	50.9	13.4	12
4-6 Teachers	45.1	20.5	9	44.6	7.0	5
Inner-City Teachers	37.8	20.3	4	43.4	12.8	5
Non Inner-City Teachers	46.0	14.7	17	50.6	11.5	14
White Teachers	45.2	16.7	17	48.3	11.9	16
Minority Teachers	41.0	11.0	4	48.5	19.1	2

BIBLIOGRAPHY

Athanases, S. (1991, April). *Alternative assessments of literacy teaching: Results of a two-year teacher assessment project.* Paper presented at the American Educational Research Association annual meeting, Chicago, Il.

Berliner, BethAnn, Mata, Susana, Zalles, Dan, and Little, Judith Warren. (1987). *Improving student teaching through clinical supervision.* **Volume two: Supervision and support through the eyes of student teachers and first year teachers.** San Francisco, CA: Far West Laboratory for Educational Research and Development.

Borko, Hilda, Lalik, Rosary, Livingston, Carol, Pecic, Kathleen, and Perry, Diana (1986). **Learning to teach in the induction year: Two case studies.** Paper presented at the annual meeting of the American Educational Research Association.

Boyer, Ernest L. (1983). **High school: A report to the Carnegie Foundation for the advancement of teaching**. New York: Harper & Row.

California State Department of Education. (1985). **Mathematics framework for California public schools, kindergarten through grade twelve.** Sacramento: California State Department of Education.

California State Department of Education. (1986). **Recommended readings in literature, K-8.** Sacramento: California State Department of Education.

California State Department of Education. (1987). **English/language arts framework for California public schools, kindergarten through grade twelve.** Sacramento, CA: Author.

California State Department of Education. (1988). **English-language arts model curriculum guide: Kindergarten through grade eight**. Sacramento: California State Department of Education.

California State Department of Education. (1990). **Science framework for California public schools, kindergarten through grade twelve.** Sacramento: California State Department of Education.

Colley, R. and Lujan, P. (1982). *A structural analysis of speeches by Native American students.* In F. Barkin, E. Brandt, & J. Ornstein-Galacia, **Bilingualism and language contact**. New York: Teacher College Press.

Delpit, L. (1986). *Skills and other dilemmas of a progressive black educator.* **Harvard Educational Review**, 56 (4), 379-385.

Delpit, L. (1988). *The silenced dialogue: Power and pedagogy in educating other peoples' children.* **Harvard Educational Review**, 58 (3), 280-298.

Estes, Gary D., Stansbury, Kendyll, and Long, Claudia. **Assessment component of the California new teacher project: First year report.** San Francisco: Far West Laboratory for Educational Research and Development, March 1990.

Estes, G., Stansbury, K., Long, C., and Wolf, K. (1991). **Assessment component of the California New Teacher Project: Second year technical report, volume I.** San Francisco, CA: Far West Laboratory for Educational Research and development, February 1991.

Foster, M. *It's cookin now: A performance analysis of the speech events of a black teacher in an urban community college.* **Language in Society**, 18 (1), 1-29.

Goodlad, John I. (1984). **A place called school: Prospects for the future**. New York: McGraw-Hill.

Grant, Carl and Zeichner, Kenneth. (1981). *Inservice support for first year teachers: The state of the scene.* **Journal of Research and Development in Education,** 14:99-111.

Griffin, G. and Millies, S. (1986). **The beginning years of teaching**. Springfield: Illinois State Board of Education.

Graves, D. (1983). **Writing: Teachers and children at work.** Portsmouth, NH: Heinemann.

Haertel, E. (1990, April). *From expert opinions to reliable scores: Psychometrics for judgement-based teacher assessment--draft only.* Paper presented at the American Educational Research Association annual meeting, Boston, MA.

Heath, S. (1983). **Ways with words**. Cambridge: Cambridge University Press.

Hollins, E. (1982). *The Marva Collins story revisited.* **Journal of Teacher Education,** 33 (1), 37-40.

Holmes Group, Inc. (1986). **Tomorrow's teachers: A report of The Holmes Group.** East Lansing, MI: The Homes Group.

Huling-Austin, Leslie. (1988). **A synthesis of research on teacher induction programs and practices**. Paper presented at the annual meeting of the American Educational Research Association.

Klein, S., & Stecher, B. (1991, January). **Final report for English/language arts tasks**. Santa Monica, CA: The RAND Corporation (AR-4005-STATE).

Kleinfeld, J. (1974). *Effective teachers of Indian and Eskimo high school students.* In J. Orvik & R. Barnhardt (Eds.), **Cultural influences in Alaskan native education.** Center for Northern Educational Research, University of Alaska, Fairbanks.

Ladson Billings, G. (1990). *Culturally relevant teaching.* **The College Board Review**, 155, 20-25.

Leinhardt, Gaea. (1989). *Math lessons: A contrast of novice and expert competence.* **Journal for Research in Mathematics Education**, 20, 52-75.

Michaels, S. and Cook-Gumperz. (1979). *A study of sharing time with first-grade students: Discourse narratives in the classroom.* **Proceedings of the fifth annual meeting of the Berkeley Linguistics Society**. Berkeley, CA: University of California.

Moir, E. (1990). *Article* in **New Teacher News** (Newsletter of CNTP), 3 (1), October 1990.

Odell, Sandra. (1986). *Induction support of new teachers: A functional approach.* **Journal of Teacher Education**, 26-29.

Quellmalz, E. (1985). *Needed: Better methods for testing higher-order thinking skills.* **Educational Researcher**, 17 (5), 5-14.

Shulman, Lee. (1987). *Knowledge and teaching: Foundations of the new reform.* **Harvard Educational Review**, 57, 1-22.

Shulman, L.S. and Sykes, G. (1986). **A national board for teaching? In search of a bold standard**. A report for the Task Force on Teaching as a Profession. New York: Carnegie Corporation.

Stiggins, R., Conklin, N., and Faires & Associates. (In press). **Classroom assessment: A task analysis**. Albany, NY: SUNY Press.

Southwest Regional Laboratory. (1990). **1988-1989 Evaluation Report**. Los Alamitos, CA: Southwest Regional Laboratory.

Southwest Regional Laboratory. (1991). **1989-1990 Evaluation Report**. Los Alamitos, CA: Southwest Regional Laboratory.

Taylor, O. and Lee, D. (1987). *Standardized tests and African-American children.* **Negro Educational Review**, 38 (2-3), 67-80.

Veenman, Simon. (1984). *Perceived problems of beginning teachers.* **Review of Educational Research**, 54, 143-178.

Ward, B. (1991). **Final report on the evaluation of the inner city new teacher retention projects**. Los Alamitos, CA: Southwest Regional Education Laboratory.

Watkins, Richard. (1985). **A practitioner review of the content validity and passing standards of the California Basic Educational Skills Test**. Sacramento: Commission on Teacher Credentialing.

Wheeler, Pat. (1986-87). *The relationship between grade six test scores and the length of the school day.* **Educational Research Quarterly**, 11 (3), 10-17.

Wheeler, Pat, Hirabayashi, J.B., Maretinson, J., and Watkins, R.W. (1988). **A study on the appropriateness of fifteen NTE specialty area tests for use in credentialing in the state of California.** Emeryville, CA: Educational Testing Service.

Wilson, Suzanne. (1988). **Understanding historical understanding: Subject matter knowledge and the teaching of teachers**. A dissertation submitted to Stanford University.

Wise, Arthur, Darling-Hammond, Linda, Berry, Barnett, Klein, Stephen P. (1987). **Licensing teachers: Design for a teacher profession.** Santa Monica, CA: The Rand Corporation.